Koordination im Deutschen und ihre syntaktische Desambiguierung

ABHANDLUNG
zur Erlangung der Doktorwürde
der
PHILOSOPHISCHEN FAKULTÄT
der
UNIVERSITÄT ZÜRICH

Vorgelegt von
Simon Clematide
Amriswil (TG)

Angenommen im Sommersemester 2006
auf Antrag von
Prof. Dr. Michael Hess

Zürich, 2009

Danksagung

Ich möchte allen Personen ganz herzlich danken, welche mich am Institut für Computerlinguistik in Zürich unterstützt und angeregt haben bei meiner Arbeit. Besonders erwähnen möchte ich (in alphabetischer Reihenfolge) Maya Bangerter, Sonja Brodersen, Alexandra Bünzli, Michael Hess, Tobias Kaufmann, Manfred Klenner, Cerstin Mahlow, Beat Rageth, Fabio Rinaldi, Gerold Schneider, Rolf Schwitter, Martin Volk.

Ein grosser Dank für ihre Unterstützung geht an meine Eltern, Geschwister, Freunde und Nachbarn. Die Geduld und den Rückhalt, den mir meine Frau und meine Kinder geschenkt haben, werde ich nie vergessen.

Zusammenfassung

Diese Arbeit untersucht mit computerlinguistischen Methoden die qualitativen und quantitativen Eigenschaften der koordinierten Strukturen auf der Ebene von Wörtern, Wortgruppen und Sätzen in der deutschen Sprache auf der empirischen Basis von syntaktisch interpretierten und annotierten Korpora (Baumbanken). Unzulänglichkeiten des Annotationsmodells und Probleme der Datenkonsistenz in der Beschreibung von Koordinationsphänomenen werden diagnostiziert und Verbesserungen vorgeschlagen.

Die Leistungsfähigkeit der automatischen Syntaxanalyse bezüglich Koordinationskonstruktionen über uneingeschränkten Texten wird für unterschiedliche sprachtechnologische Ansätze (Chunking, Dependenz- und Phrasenstruktur-Parsing) evaluiert.

Für eine Optimierung der syntaktischen Analyse von koordinierten Strukturen werden Merkmale wie Kopfdistanz, morphologische und semantische Ähnlichkeit erhoben sowie Experimente zur automatischen Klassifikation von Kommas hinsichtlich ihrer koordinativen Funktion gemacht.

Die syntagmatische Beziehung zwischen Köpfen von Konjunkten wird mit typischen lexikalisch-semantischen Relationen verglichen und die automatische Extraktion von Konjunktköpfen zur Akquirierung und Ergänzung semantischer Ressourcen getestet.

Für die Auswertungen wurde in der logischen Programmiersprache PROLOG ein einfaches, aber flexibles Repräsentationsformat für Syntaxgraphen und eine Software-Bibliothek entwickelt, welche bezüglich der Suche, Extraktion und Transformation syntaktischer Strukturen unterschiedlicher Ausprägung eine einheitliche Programmierschnittstelle bietet.

Abstract

This thesis explores the qualitative and quantitative properties of coordinated structures in the German language. It does so at the level of words, word groups, and clauses, using computational linguistics methods on the empirical basis of syntactically interpreted corpora, so-called tree banks. We locate shortcomings of the annotation model, discuss data inconsistencies in the description of coordination phenomena and suggest improvements.

Furthermore, we evaluate the performance of the automatic analysis of coordinated constructions over unrestricted texts by applying different language technological approaches such as chunking, dependency parsing and phrase-structure-based parsing.

To optimize the syntactic analysis of coordinate structures, we study features such as head distance, morphological and semantic similarity, and present results on machine-learning experiments where commas had to be classified according to whether they act as coordinating conjunctions or not. This is followed by a

comparison of syntagmatic relation between heads of coordinated structures and lexical-semantic relations typically found in a thesaurus. Further, we test automatic acquisition and extension of semantic resources by using the extracted heads.

For our evaluations, we developed a software library, using the logic programming language PROLOG, which is based on a simple, yet flexible representation format for syntax graphs. This offers a uniform application interface for searching, extracting, and transforming linguistic structures that originate from different sources.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Kapitel 1

Einleitung

1.1 Baumbanken für Deutsch

Während für das Englische seit Anfang der 90er-Jahre grössere syntaktisch annotierte Ressourcen, welche manuell validiert sind, existieren, ist für Deutsch mit dem NEGRA-Korpus erst seit 1999 eine solche Ressource für schriftliche Sprache verfügbar.

1.1.1 Die Ziele der NEGRA-Baumbank

Das NEGRA-Projekt (Skut u. a. 1997) bezieht für seine Baumbank die grundlegende Ausrichtung des Annotationsschemas von den wichtigen Baumbanken für das Englische: der Penn-Treebank (Marcus u. a. 1993) sowie dem SUSANNE-Korpus (Sampson 1995):

- Deskriptive Adäquatheit: Die Phänomene der Sprache sollen nicht erklärt werden durch die Annotation, sondern strukturell beschrieben.
- Theorie-Unabhängigkeit: Auch wenn für die Beschreibung von Sprache immer Theorie und Begrifflichkeit notwendig sind, soll die Annotation selbst möglich wenig theoriespezifische Annahmen treffen.
- Empirisches Primat: Das Annotationsschema orientiert sich an den Bedürfnissen, welche durch die schriftliche und mündliche Sprachwirklichkeit und ihre möglichst vollständig abdeckende Beschreibung gegeben sind, und nicht an theorie-immanenten Kriterien.

Eine stark konfigurationale Sprache wie Englisch mit strikter Wortstellung und Abhängigkeiten, welche relativ natürlich mit den Mitteln der Konstituenz ausgedrückt werden können, lässt sich recht gut erfassen mit folgenden Beschreibungsmitteln:

- Kontextfreie Phrasenstruktur (*context-free backbone*): Die primäre Struktur ist die Baumstruktur, welche insbesondere die Argument-Struktur enkodiert.

Nicht-lokale Abhängigkeiten werden darüber hinaus durch leere Elemente und ihre Koindizierung ausgedrückt.

- Syntaktische Kategorien: Die Konstituenten werden auf Grund ihres lexikalischen Kopfes kategorisiert.
- Grammatische Funktionen: Abhängigkeiten und Modifikationsverhältnisse werden soweit möglich bzw. nötig durch verfeinerte Knoten-Bezeichnungen ausgedrückt.

Während sich die meisten linguistischen Theorien bezüglich angenommener syntaktischer Kategorien und grammatischer Funktionen nur wenig unterscheiden, gibt es bezüglich phrasenstrukturellem Aufbau zwischen verschiedenen Theorien, aber auch schon zwischen unterschiedlichen Versionen eines gemeinsamen Theorieansatzes oft eine starke Varianz, welche sich nur schwer mit der Forderung nach Theorie-Unabhängigkeit vereinen lässt.

Für Sprachen mit freier Wortstellung und einer grösseren Anzahl diskontinuierlicher Erscheinungen bedeutet das Aufrechterhalten einer kontextfreien phrasenstrukturellen Beschreibungsebene eine aufwendige Abbildung auf eine theoretisch zugrundeliegende konfigurationale Grundstruktur.

Im NEGRA-Projekt wird ein solcher Ansatz für die Grundannotation verworfen, und man versucht stattdessen, die strukturellen und funktionalen Zusammenhänge ohne leere Elemente und Koindizierung zu beschreiben. Ein solches Unterfangen ist mit Bäumen mit überkreuzenden Kanten machbar.

1.1.2 Umgang mit Annotations-Fehlern in annotierten Korpora

Trotz aller Bemühungen um Sorgfalt und Konsistenz enthalten Baumbanken Fehler auf allen Ebenen der Aufbereitung und Annotation: Fehler in der Wortsegmentierung (Tokenisier-Fehler), der Wortartenklassifikation (Part-Of-Speech-Tagging-Fehler) und am häufigsten in der syntaktischen Struktur (Syntax-Fehler). Eine Möglichkeit zur Qualitätssicherung besteht darin, die intellektuelle Annotation von 2 verschiedenen Personen erstellen zu lassen. Falls die beiden Annotationen identisch sind, werden sie als korrekt erachtet. Falls Differenzen bestehen, d.h. kein *Inter-Annotator Agreement* entsteht, muss die Annotation revidiert werden.

Für das NEGRA-Korpus gibt Brants (2000a) folgende Zahlen zur Annotationsübereinstimmung[1]: Auf der Ebene der STTS-Wortartenbestimmung ist das F-Mass der Übereinstimmung zwischen 2 Annotatoren bei 98.57% (bei insgesamt 147212 Token). Bei 3 Wortarten-Tags pro 200 Token gibt es im Schnitt keine Übereinstimmung. Die Übereinstimmung mit der Schlussannotation nach der Bereinigung ist bei 98.8%. Die Annotationsübereinstimmung für die syntaktischen Strukturen ist 93.7% für die reine Phrasenstruktur ohne Beschriftung, 92.4% mit Beschriftung, sowie 88.5% wenn die syntaktische Funktion ebenfalls berücksichtigt wird. Bezogen auf ganze Sätze bedeuten diese Zahlen, dass etwa 52% (Brants u. a. 2003, 82)

[1] Entsprechende Werte für die Penn-Treebank finden sich in Marcus u. a. (1993).

aller Sätze die identische Annotation erhalten vor der Revision – die reine Phrasenstruktur ohne Beschriftung stimmt in 68% der Sätze überein. Diese Zahlen verdeutlichen die Notwendigkeit von Mehrfachannotation für konsistente Resultate. Allerdings können auch übereinstimmende Annotationen übereinstimmend falsch sein. Die Verwendung von halbautomatischen Verfahren, welche etwa 71% (Brants u. a. 2003, 80) der Phrasen korrekt vorschlagen, können mit ihren Vorschlägen die Annotatoren auch konsistent zu falschen Parsebäumen verleiten.

Die Qualität von Annotationshandbüchern mit klaren Strukturvorgaben und operationalisierbaren Tests sowie das Vermeiden von Klassen mit hoher Verwechslungsgefahr sind dabei entscheidend. Auf der Ebene der Wortartenbestimmung mit dem STTS-Tagset ist die Unterscheidung von Eigennamen (NE), Substantiven (NN) und fremdsprachlichem Material (FM) eine notorische Verwechslungsquelle. Auf der Ebene der Phrasenstruktur sind besonders seltenere Konstruktionen schlecht interannotiert. Nach Brants (2000a) sind gerade die seltenen Koordinationsstrukturen diejenigen, mit den schlechtesten Übereinstimmungen: Koordinierter Satzkomplementierer (CCP) 50%[2], Koordination ungleicher Kategorien (CO) 56%.

Neben der Qualitätssicherung durch intellektuellen Abgleich und Korrektur, was nach Brants (2000a) im Schnitt bis etwa 8 von 10 Minuten Arbeitszeit pro Satz erzeugt, wurden verschiedene Verfahren für die automatische Konsistenzsicherung bzw. Fehlerentdeckung in der Literatur vorgeschlagen. Dickinson (2005) präsentiert etwa einen Ansatz, der mit einfachen N-Grammen die Konsistenz ermittelt. Wallis (2003) schlägt ein Verfahren vor, wo konstruktionsspezifisch über einer Baumbank korrigiert wird.

Ein wesentlicher deskriptiver Teil dieser Arbeit besteht ebenfalls darin, die Annotationen der Konstruktion „Koordination“ in Baumbanken auszuwerten. Durch die systematische Auswertung ist einiges an Fehlern, Inkonsistenzen und Problem-Annotationen sichtbar geworden. Da die Korpus-Erstellung nicht zum Umfang dieses Projekts gehört, stellt sich die Frage, wie mit diesen Fällen umgegangen werden soll. Grundsätzlich ergeben sich folgende Möglichkeiten:

- Ausschluss: Problemsätze werden sobald als solche identifiziert von der Untersuchung ausgeschlossen.
- Korrektur: Einfache offensichtliche Fehler können beseitigt werden, insbesondere, wenn sie auf der Ebene der Tokenisierung und Wortartenbestimmung vorliegen. Fälle, wo nur eine Beschriftung einer Phrase oder Funktionsbezeichnung falsch sind, fallen ebenfalls darunter.
- Ignorieren: Die Fehler werden als „Rauschen“ in den Daten akzeptiert.

Für die wortbasierten Auswertungen wurden die offensichtlichen Annotationsfehler korrigiert, ebenso die Fälle, welche reine Beschriftungsprobleme in der Syntaxstruktur waren. Komplexe syntaktische Umstrukturierung waren nicht möglich –

[2] Von dieser Kategorie verbleibt in NEGRA letztlich nur ein Exemplar.

insbesondere da die Grenzlinie zwischen „Fehlannotation“ und unterschiedlicher Interpretation der knapp gehaltenen Annotationsrichtlinien eine Unschärfe zeigt. Auch wenn diese Probleme für die quantitativen Auswertungen ignoriert werden, bieten die qualitativen Beurteilungen in Beispielsätzen Gelegenheit, die Problemfälle, welche oft hinter „Fehlern“ stecken, zu diskutieren.

Da die Fehler oft erst während der qualitativen Exploration der quantitativen Auswertungen zutage kommen, stellte sich das Problem der Konsistenz der Auswertungen über den verschiedenen Korrektur-Versionen des Korpus. Die meisten quantitativen Auswertungen in Tabellen und Listenform in dieser Arbeit werden deshalb automatisch mit selbstentwickelten Aufbereitungswerkzeugen aus den aktuellsten Daten erzeugt, indem die notwendigen Generierungsschritte mitsamt ihren Vorbedinungen und Abhängigkeiten über Konfigurationsdateien des Software-Entwicklungswerkzeugs `make`[3] verwaltet werden.

1.1.3 Weitere benutzte Baumbanken

Neben dem Hauptkorpus NEGRA werden noch zwei weitere im ähnlichen Stil annotierte Ressourcen in die Diskussion einbezogen: Das TIGER-Korpus Version 1 (Brants u. a. 2002), welches in vielen Aspekten eine Fortführung der NEGRA-Annotationstradition darstellt und eine willkommene Ausdehnung der gut 20602 NEGRA-Sätze um 40020 TIGER-Sätze bedeutet. Allerdings wurde gerade im Bereich der Satzkoordination im TIGER-Korpus eine Ergänzung des Annotationsmodells vorgenommen. Dabei wird motiviert durch Koordinationsphänomene wie in Abbildung 1.1 auf der nächsten Seite, mit Hilfe von sekundären Kanten die mangelnde strukturelle Annotation von koordinierten finiten Prädikatsteilen kompensiert.

Da sowohl das NEGRA-Korpus wie das TIGER-Korpus auf Zeitungstexten der „Frankfurter Rundschau“ (NEGRA auf Texten aus dem Jahr 1992, TIGER auf Texten aus 1995 und 1997) basiert, wurde für gewisse Auswertungen ein kleineres, 3000 Sätze umfassendes Korpus (CZ-Korpus) beigezogen, welches im Rahmen eines Habilitationsprojekts (Volk 2001) aus Texten der „ComputerZeitung“ in Zürich in der Tradition von NEGRA annotiert worden ist.

1.1.4 Notationskonventionen

Im Fliesstext werden alle Kürzel (engl. *tags*) für Wortarten (siehe Anhang A.1 auf Seite 297 für eine vollständige Liste mit Kurzbeschreibungen), Phrasenkategorien (siehe Anhang A.2 auf Seite 299) und grammatischen Funktionen (siehe A.3 auf Seite 300) aus den von mir untersuchten Korpora in serifenloser Schrift gesetzt. Die Serifenvariante verwende ich, wenn in einem allgemeinen linguistischen Sinn von diesen Kategorien die Rede ist. In Tabellen oder Beispielsätzen verwende ich

[3] Siehe Dalheimer (1998) für eine kompakte Beschreibung und `http://www.gnu.org/software/make` für die Software.

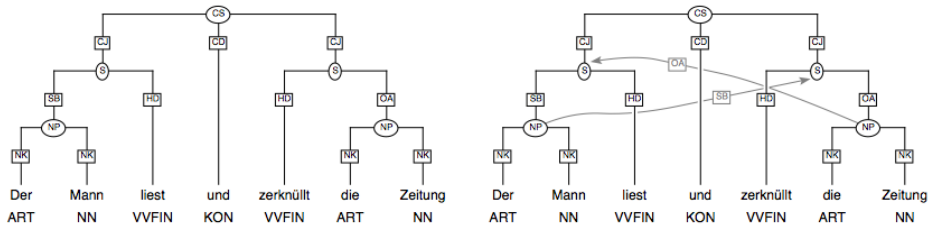


Abbildung 1.1: Annotation von koordinierten Verbalteilen in NEGRA (Brants u. a. 1999, 88) vs. TIGER (Albert u. a. 2003, 118)

durchgehend die Serifenschrift. Abkürzungen wie PP für Präpositionalphrase verwende ich sowohl für Singular- wie Pluralverwendungen und verzichte auf Formen wie PPen oder PPs.

Die Quellen der Beispielsätze sind jeweil in eckigen Klammern angegeben. So verweist etwa $[N_{168}]$ auf Satz 168 des NEGRA-Korpus.

1.1.4.1 Baumdiagramme und Kastendiagramme

Baumdiagramme für Phrasenstrukturen sind eine gängige Visualisierung der Teil-Ganzes-Beziehung und der linearen Präzedenz-Beziehung, welche durch die Konstituentenanalyse geschaffen wird. Eine gängige äquivalente Notation dazu sind die sogenannten Kastendiagramme. Wenn mit funktionalen Klassifikationen von Phrasen gearbeitet wird wie in NEGRA, braucht es zusätzliche Notationsebenen. Für die Baumdiagramme hat sich beeinflusst durch das zur Korpusannotation von NEGRA verwendete graphische Werkzeug `annotate` eine Tradition herausgebildet, welche wie in Abbildung 1.2 die Phrasenbeschriftungen in Ellipsen und die Funktionsbeschreibungen in Rechtecken notiert.

Beide Darstellungen haben ihre Vorzüge: Kastendiagramme betonen stärker, wo untergeordnete Konstituenten eingehängt werden. Insbesondere bei grösseren Strukturen erschliesst sich der wachsende Aufbau in der Vertikalen dem Auge schneller, während die dünnen Verbindungslinien der Baumdiagramme die Konstituenz weniger direkt wiedergeben. Bei den Kastendiagrammen hingegen stören isolierte Elemente wie die Interpunktion, welche nicht in die Konstituentenstruktur eingebaut sind, als Unterbrechungen der horizontalen Kasten etwas stärker als bei den Baumdiagrammen.

Aus Platzgründen werden in den Beispielen oft keine Diagramme verwendet, sondern die relevanten (und nur die relevanten) strukturellen Informationen in der Schreibweise der eckigen Klammern angegeben, wie es in Abbildung 1.4 auf der nächsten Seite gezeigt wird.

Die Darstellung von diskontinuierlichen Strukturen, d.h. Konstituenten, welche unterbrochen werden von Material, das zu andern Konstituenten gehört, ergibt zusätzliche Anforderungen. In der Baumdiagramm-Darstellung zeigen sich kreuzen-

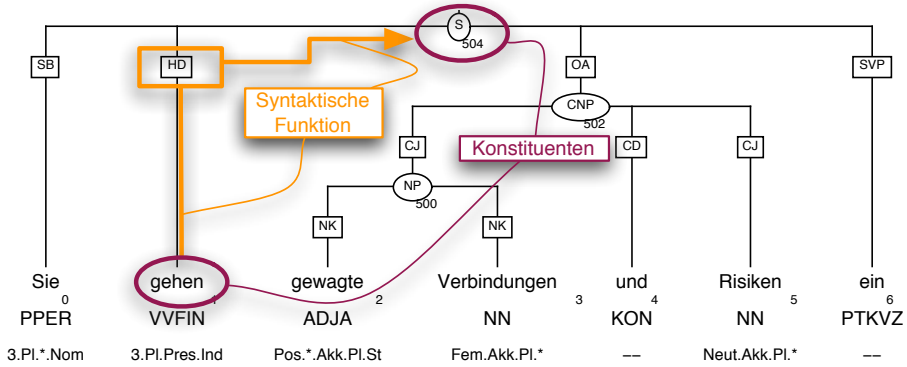


Abbildung 1.2: Konstituenten in Baumdiagramm-Form wie sie im NEGRA-Annotationswerkzeug und TIGERSearch-Werkzeug dargestellt werden. Sowohl die syntaktischen Nicht-Terminalknoten in oval als auch die Terminalknoten sind Konstituenten. Die Kategorien der Terminale (Wortart) werden graphisch anders behandelt als die Kategorien der Nicht-Terminale (Phrasenbeschriftung), obwohl sie dieselbe Funktion erfüllen.

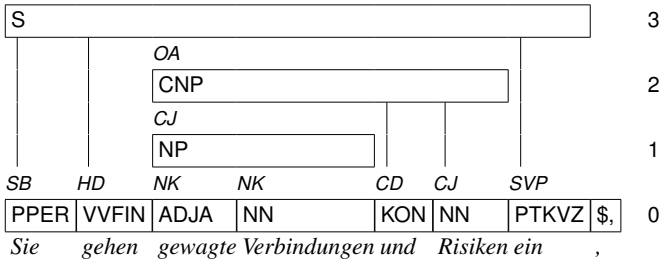


Abbildung 1.3: Konstituenten im Kastendiagramm. Die Kategorien der Terminale (Wortart) und Nicht-Terminale (Phrasenbeschriftung) werden konsistent als „Überbau“ behandelt. Die Höhe der syntaktischen Ebenen wird auf der rechten Seite durchnummeriert.

Sie gehen $[_{NP-CJ}$ gewagte Verbindungen] $[_{KON-CD}$ und] $[_{NN-CJ}$ Risiken]] ein

Abbildung 1.4: Partielle Annotation mittels eckiger Klammern und Subskripten der Kategorie und Funktion nach dem Schema $[_{CAT-FUN} \ldots]$.

de Kanten wie in der Abbildung 1.5 auf der nächsten Seite, welche der Darstellung im TIGERSearch-Werkzeug entspricht. Die entsprechende Darstellung als Kastendiagramm enthält die Abbildung 1.6 auf der nächsten Seite. Diskontinuierliche Konstituenten zeigen sich dabei als Balkenkasten mit fehlenden „Seitenwänden“ und dazwischen durchlaufenden Dominanzlinien.

Im Fall von diskontinuierlichen Strukturen ist die textuelle Kurznotation mit eckigen Klammern nicht mehr besonders augenfällig. Aus Platzgründen wird trotzdem manchmal diese Darstellungsform gewählt – die eingeschobenen Konstituententeile sind dann „ausgeklammert“ mittels $\lfloor \ldots \rfloor$ wie in Abbildung 1.7 auf der nächsten Seite.

Eine formale und detaillierte Darstellung der strukturellen Repräsentation von syntaktischen Beschreibungen findet sich im Kapitel 6 auf Seite 261.

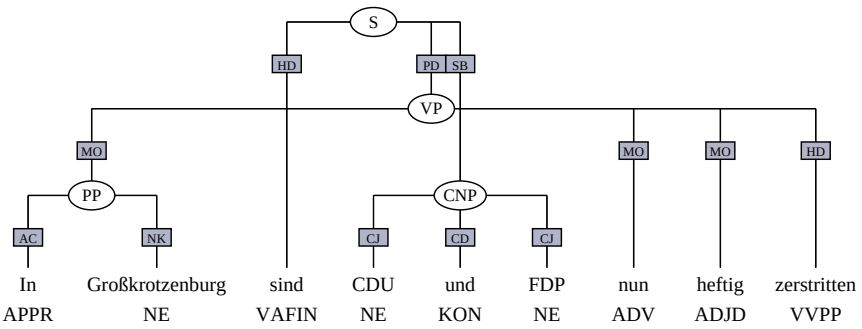


Abbildung 1.5: Darstellung der Diskontinuierlichkeit mittels überkreuzender Kanten in der Baumdarstellung.

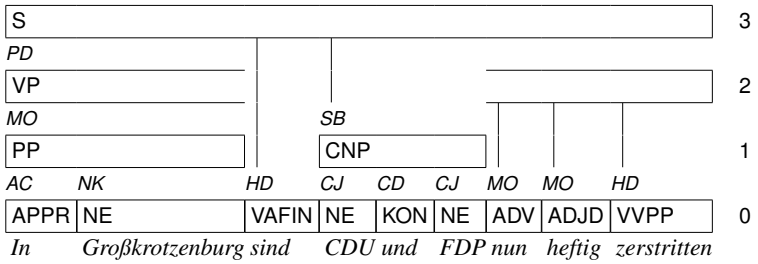


Abbildung 1.6: Darstellung der Diskontinuierlichkeit in Kastendiagrammen. Diskontinuierliche Konstituenten haben offene Seiten, zwischen denen Dominanzlinien durchlaufen.

[$_{VP}$ In Großkrotzenburg ⌊ [$_{VAFIN-HD}$ sind] ⌋ ⌊ [$_{CNP}$ CDU und FDP] ⌋ nun heftig zerstritten]

Abbildung 1.7: Partielle Annotation von diskontinuierlichen Strukturen mittels eckigen Klammern und Aussparungsmarkierungen ⌊...⌋.

Kapitel 2

Linguistisches Phänomen Koordination

There is no such thing as a theory-free description.

2.1 Einleitung

Koordination ist ein syntaktisches Phänomen, das im Deutschen auf verschiedensten Stufen sprachlicher Realisation vorkommt. Eine einfache beschreibende Einteilung nach Zifonun u. a. (1997) geht von 4 Stufen aus und ist im Folgenden mit prototypischen und grammatisch einfachsten Beispielen aus NEGRA illustriert:

Morphemkoordination/Teilwortkoordination: Verknüpfen von Teilwörtern[1]:

(1) [...] das Gerichtswesen wird jetzt selbständig und von den übrigen *Polizei- und Verwaltungsarbeiten* getrennt. [N_{168}]

Wortkoordination: Verknüpfen von selbstständigen Wörtern:

(2) Selbst die flotteren Passagen werden nie *ausgelassen und fröhlich*. [N_{35}]

Wortgruppenkoordination: Verknüpfen von Einheiten, welche aus mehr als einem Wort bestehen:

(3) Oder saß es wieder einmal immer nur mit der Teeflasche *im Auto und in der Sportkarre*? [N_{9328}]

Satzkoordination: Verknüpfen von vollständigen Sätzen

[1] Siehe Luschützky (2000) für einen Überblick zum Begriff „Morphem“. Hier bezeichnet Morphemkoordination einfach die Verknüpfung von mindestens einer Einheit, welche kleiner als ein vollständiges Wort ist.

(4) Wenn es eine Organisation gibt, der vertraut wird und die etwas erreichen kann, ist das die Kirche. [N_{11978}]

Obige Einteilung idealisiert einerseits die sprachliche Wirklichkeit, indem viele koordinierte Strukturen durch dieses Schema nicht richtig erfasst werden, da Verknüpfungen zwischen unterschiedlichen Stufen vorkommen. Anderseits ist sie durch ihre allgemeine linguistische Begrifflichkeit zu unpräzise, um grammatikalisch prägnante Aussagen über Möglichkeiten und Grenzen der Koordination zu machen.

Das erste Problem lässt sich selbstverständlich durch den Beizug von breiter abdeckenden, beschreibenden Grammatiken angehen. Wichtig ist mir jedoch, die Diskussion mit empirischen Ergebnissen aus der Untersuchung der syntaktisch annotierten Korpora zu unterfüttern, welche möglichst präzise und trotzdem übersichtlich dargestellt werden sollen.

Das zweite Problem lässt sich mit einem ausgebildeteren Theorieansatz eindämmen; es beinhaltet aber auf jeden Fall eine Diskussion zur Behandlung von Ersparungen (oft auch mit dem Begriff „Ellipsen“ bezeichnet), welche im Zusammenhang mit koordinierten Strukturen eine prominente Funktion einnehmen. Die Auffassungen darüber, welche Strukturen koordiniert werden können, hängen insbesondere davon ab, ob die Analysen dezidiert oberflächennah operieren, d.h. keine oder möglichst wenig unsichtbare Elemente postulieren, oder ob sich die syntaktischen Analysen stärker an der semantischen Interpretation der Strukturen orientieren. Diese beiden Haltungen schliessen sich nicht aus, wie man etwa vermuten könnte, denn gerade die Annotationsphilosophie von NEGRA zeigt dies. Die annotierten syntaktischen Strukturen im NEGRA sind immer auf eine Interpretation bezogen, welche auch Ergänzungen machen muss, diese aber nicht immer explizit annotiert[2]. Allerdings leidet unter einem solchen Ansatz die Generalisierbarkeit der in den annotierten Strukturen implizit gegebenen Grammatik, solange die mitinterpretierten Strukturen nicht explizit annotiert werden (vgl. Abbildung 1.1 auf Seite 5).

Eine differenzierendere Theorie wird zudem Koordinationsphänomene von ähnlichen, aber in ihrer Funktion doch zu unterscheidenden Phänomenen wie der Apposition und Parenthese trennen

Eine gute Zusammenstellung wichtiger linguistischer Literatur zur Koordination findet sich in Zifonun u. a. (1997, 2360), welche insbesondere auch van Oirsouw (1993) erwähnt, der einen historischen Überblick über die Ansätze zur Koordination in der theoretischen Linguistik seit Chomskys „Syntactic Structures“ (1957) gibt. Mit Johannessen (1998) liegt ein breite komparative Analyse im theoretischen Rahmen des minimalistischen Programmes der Generativen Grammatik vor. Eine Abhandlung im gleichen Theorierahmen machen te Velde (2005) und Camacho (2003). Breitere Analysen zu Koordinationsphänomenen in der Kategorialgrammatik gibt Houtman (1994) und Steedman (2002). Das Thema „Koordinationsel-

[2]Die Verwendung von sekundären Kanten zur Integration von mitinterpretiertem Material ist in TIGER gemacht, sollte aber ursprünglich schon im NEGRA-Korpus realisiert werden.

lipsen“ wird ausgiebig besprochen in Klein (1993), weitere Entwicklungen sind in Schwabe und Zhang (2000) präsentiert.

2.1.1 Begriffe zur Beschreibung von Koordinationsphänomenen

Um die Diskussion der folgenden Abschnitte präzis halten zu können, kläre ich im folgenden die Verwendung einiger wichtiger Fachbegriffe. Ich halte mich mehrheitlich an Zifonun u. a. (1997, 2362f.), verwende jedoch die Unterscheidung zwischen syndetisch, asyndetisch und monosyndetisch des Grammatik-Dudens (Dudenredaktion 2005, §1408).

Konjunktoren (auch „nebenordnende Konjunktionen“ genannt im Duden) sind ein- oder mehrteilige Ausdrücke, welche zwei sprachliche Einheiten mit verträglicher syntaktischer Funktion gleichrangig verbinden.

Juxtaposition meint das gleichrangige Verbinden von zwei oder mehr sprachlichen Einheiten ohne Verwendung eines lexikalischen Konjunktors. Mündlich wird dies durch progredientes Tonmuster, schriftlich durch das Setzen von Interpunktion wie Komma, Semikolon, Gedankenstrich oder Doppelpunkt ausgedrückt.

Konjunkt (auch „Koordinaten“ genannt in Zifonun u. a. (1997)) meint die sprachlichen Einheiten, welche durch Konjunktoren oder Juxtaposition miteinander verbunden werden.

Syndetisch koordinierte Struktur meint das Gefüge, das nur aus mit Konjunktoren verbundenen Konjunkten besteht.

Asyndetisch koordinierte Struktur meint das Gefüge, das aus Konjunkten besteht, welche in Juxtaposition stehen.

Monosyndetisch koordinierte Struktur meint das Gefüge, wo das letzte Konjunkt mit einem Konjunktor verbunden ist und alle vorangehenden durch Juxtaposition – es braucht dabei mindestens 3 Konjunkte.

Katalepse (auch „Rückwärtsellipse“) bezeichnet eine Auslassung, deren Füllung im nachfolgenden Text erscheint, bzw. nachfolgend verbalisiert wird.[3]

Analepse (auch „Vorwärtsellipse“) bezeichnet eine Auslassung, deren Füllung im vorangehenden Text erschienen ist, bzw. vorangehend verbalisiert wurde.

[3] Ob Auslassungen syntaktisch repräsentiert werden oder ob sie nur in der semantischen Interpretation berücksichtigt bzw. rekonstruiert werden, ist ausgesprochen theorieabhängig und kontrovers. Siehe die weiterführende Diskussion in Klein (1993).

2.2 Koordinierte Strukturen

Know your data.

Was wird koordiniert? In der folgenden beschreibenden Diskussion koordinierter Strukturen halte ich mich bezüglich der Reihenfolge an die Strukturklassen von Zifonun u. a. (1997, 2360), welche Koordination auf der Stufe von Morphemen, Wörtern, Wortgruppen, Phrasen sowie ganzen Sätzen beschreibt.

Vor der detaillierten Diskussion der verschiedenen Arten von Koordinationen soll ein Überblick über das Vorkommen und die Verteilung in syntaktisch annotierten Korpora gegeben werden.

2.2.1 Koordinationen in NEGRA, TIGER und CZ

Die Tabellenzusammenstellung 2.1 auf der nächsten Seite zeigt auf, wieviele koordinierte Strukturen von welcher Kategorie überhaupt auftreten. Mit mehr als 1/2 aller Vorkommen sind koordinierte Nominalphrasen (CNP) erwartungsgemäss die wichtigste Kategorie. 1/4 der Vorkommen wird von koordinierten Sätzen (CS) gestellt. Die koordinierten Adjektive (CAP) stellen nur noch 1/10, koordinierte nichtfinite Verbalphrasen (CVP) und koordinierte PP (CPP) noch rund 1/20. Es zeigt sich diesbezüglich über den 3 Korpora aus Zeitungstexten ein einheitliches Bild.

Obwohl die Häufigkeit von koordinierten Strukturen einer bestimmten Kategorie zusammenhängt mit dem Vorkommen der Kategorie überhaupt, gibt es doch starke kategorienspezifische Unterschiede zum Anteil der koordinierten Strukturen. So gibt es in NEGRA zu den 35500 PP weniger als 500 koordinierte Exemplare, was mit einem Anteil von 1/71 ein sehr niedriger Wert ist. Wie man der Tabellenzusammenstellung 2.2 auf Seite 14 entnehmen kann, gibt es in NEGRA hingegen zu den 13500 VP mehr als 500 Exemplare, was einem Anteil von über 1/27 entspricht.

Wenn man sich für die exakten relativen Anteile pro Kategorie interessiert, ergeben sich in NEGRA die Werte in Auflistung (5), welche nach dem grössten relativen Anteil der koordinierten Phrasen geordnet sind:

(5) Relative Anteile der häufigsten Koordinationskategorien in NEGRA:
CAP (13.3%), AP (86.7%); CNP (10.9%), NP (89.1%); CS (8.1%),
S (91.9%); CVP (4.0%), VP (96.0%); CPP (1.3%), PP (98.7%)

Diese Zahlen unterscheiden sich für TIGER nur unwesentlich:

(6) Relative Anteile der häufigsten Koordinationskategorien in TIGER:
CAP (13.4%), AP (86.6%); CNP (10.2%), NP (89.8%); CS (7.5%),
S (92.5%); CVP (4.3%), VP (95.7%); CPP (1.4%), PP (98.6%)

Etwas stärkere Unterschiede ergeben sich im kleinen CZ-Korpus wie die Auflistung (7) aufweist, insbesondere die erhöhten CAP-Werte stechen heraus.

NEGRA (total 9942)

in %	Anzahl	Kategorie	kumulativ
52.0	5169	CNP	52
25.7	2555	CS	78
9.1	900	CAP	87
5.5	547	CVP	92
4.8	477	CPP	97
1.7	166	CO	99
1.0	95	CAVP	100
0.3	26	CAC	100
0.1	5	CVZ	100
0.0	1	CCP	100

TIGER (total 19353)

in %	Anzahl	Kategorie	kumulativ
51.1	9894	CNP	51
24.3	4702	CS	75
9.8	1894	CAP	85
6.6	1283	CVP	92
5.4	1051	CPP	97
1.6	309	CO	99
0.9	177	CAVP	100
0.1	21	CAC	100
0.1	20	CVZ	100
0.0	1	CCP	100

CZ (total 1976)

in %	Anzahl	Kategorie	kumulativ
57.8	1141	CNP	58
19.3	382	CS	77
12.1	238	CAP	89
4.9	97	CVP	94
4.4	87	CPP	98
0.8	15	CO	99
0.5	9	CAVP	100
0.2	3	CVZ	100
0.2	3	CAC	100

Tabelle 2.1: Verteilung der koordinierten Phrasen in NEGRA, TIGER und CZ

NEGRA

Typ	Anzahl	in %	Kategorie	Anzahl	in %
(C)NP	47455	35.1	NP	42286	31.3
			CNP	5169	3.8
(C)PP	35998	26.6	PP	35521	26.3
			CPP	477	0.4
(C)S	31532	23.3	S	28977	21.4
			CS	2555	1.9
(C)VP	13523	10.0	VP	12976	9.6
			CVP	547	0.4
(C)AP	6763	5.0	AP	5863	4.3
			CAP	900	0.7

TIGER

Typ	Anzahl	in %	Kategorie	Anzahl	in %
(C)NP	96810	34.9	NP	86916	31.3
			CNP	9894	3.6
(C)PP	73780	26.6	PP	72729	26.2
			CPP	1051	0.4
(C)S	62701	22.6	S	57999	20.9
			CS	4702	1.7
(C)VP	30061	10.8	VP	28778	10.4
			CVP	1283	0.5
(C)AP	14180	5.1	AP	12286	4.4
			CAP	1894	0.7

CZ

Typ	Anzahl	in %	Kategorie	Anzahl	in %
(C)NP	8415	33.9	NP	7274	29.3
			CNP	1141	4.6
(C)PP	7976	32.1	PP	7889	31.8
			CPP	87	0.4
(C)S	5017	20.2	S	4635	18.7
			CS	382	1.5
(C)VP	2318	9.3	VP	2221	8.9
			CVP	97	0.4
(C)AP	1111	4.5	AP	873	3.5
			CAP	238	1.0

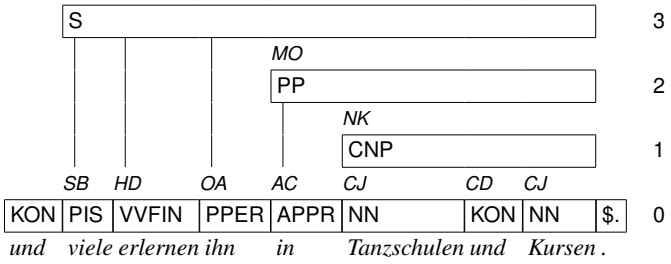
Tabelle 2.2: Verhältnis der häufigsten koordinierten Phrasen zu den unkoordinierten in NEGRA, TIGER und CZ

(7) Relative Anteile der häufigsten Koordinationskategorien in CZ:
CAP (21.4%), AP (78.6%); CNP (13.6%), NP (86.4%); CS (7.6%),
S (92.4%); CVP (4.2%), VP (95.8%); CPP (1.1%), PP (98.9%)

Unäre Phrasen Die oben gezeigten Auswertungen verwenden die Phrasen, wie sie in den Korpora „wörtlich“ annotiert sind. Das zugrunde liegenden Annotationsmodell verbietet syntaktische Kategorien, welche nur aus einer Tochterkonstituente bestehen und sogenannte „unäre“ Phrasen bilden würden[4]. Deshalb bestehen die NP, welche in der Tabellenzusammenstellung 2.2 auf der vorherigen Seite mit CNP kontrastiert sind, also immer aus mehr als einem Wort.

Das pronominale Subjekt und Objekt in Beispiel (8) ergibt auf der syntaktischen Ebene keine explizit annotierte NP, sondern die Pronomen werden unmittelbar vom S-Knoten dominiert.

(8)



[N_{3473}]

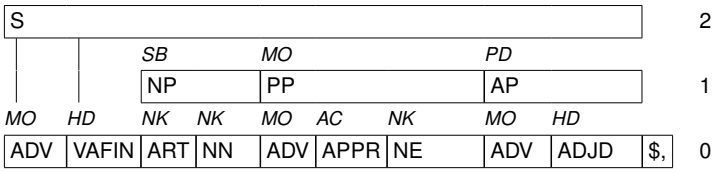
Flache Einbettung Eine weitere Eigenheit des Annotationsmodells stellt die flache Einbettung der Nominalphrasen in die PP dar, welche nicht bloss die Nominalphrasen aus einem Wort wie „Frankfurt“ in Beispiel (9a) auf der nächsten Seite betrifft, sondern auch mehrteilige Nominalphrasen mit Begleiter, Adjektiven[5] und Kernnomen wie in (9b). Im Fall des flach in die PP eingebetteten Adjektivs „phantasievollen“ in (9b) ergeben sich auszählungsrelevante Strukturunterschiede, wenn die asyndetisch koordinierte CAP gleichermassen annotiert wäre wie in (9c). Noch stärker wäre der Effekt, wenn Nominalphrasen entsprechend einer Annotationsstrategie wie in (9d) behandelt würden, welche die eingebettete NP als eigenständigen Knoten in die PP eingebaut. Dies wird beispielsweise in der Tübinger Baumbank (Müller 2004, 16) so gemacht. Im Verlauf dieser Arbeit werde ich insbesondere bei den Nominalphrasen[6] und Adjektiven diese Sachverhalte noch genauer betrachten.

[4] Strukturelle Sparsamkeit wird auch von (Bresnan 2001, 91) mit dem Prinzip der Ausdrucksökonomie angestrebt: „All syntactic phrase structure nodes are optional and are not used unless required by independent principles (completeness, coherence, semantic expressivity).“

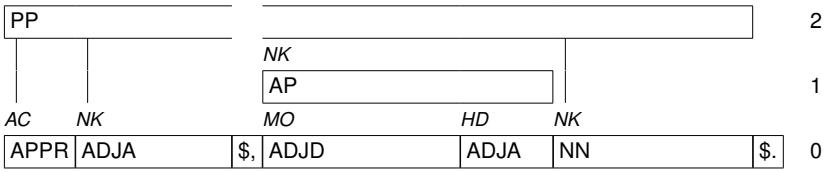
[5] Dieses Beispiel illustriert zudem, dass gerade in NEGRA asyndetische Adjektivkoordinationen nicht-koordinativ annotiert sein können. Dies wird in der Diskussion der Tabelle 2.8 auf Seite 23 noch genauer untersucht.

[6] Auch bei den Präpositionalphrasen kann man unäre Konstruktionen in Betracht ziehen, nämlich bei den Pronominaladverbien.

(9) a.



b.



[N_{997}]

c. [$_{PP}$ in [$_{CAP-NK}$ [$_{ADJA-CJ}$ phantasievollen , [$_{AP-CJ}$ abwechslungsreich gebauten]] Energiesparhäusern] .

d. [$_{PP}$ in [$_{NP}$ [$_{CAP-NK}$ [$_{ADJA-CJ}$ phantasievollen , [$_{AP-CJ}$ abwechslungsreich gebauten]] Energiesparhäusern]] .

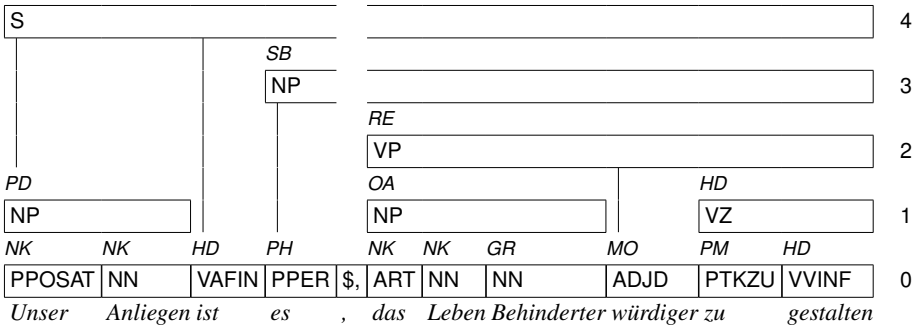
2.2.1.1 Implizite Einzelwort-NP

In diesem Abschnitt soll abgeschätzt werden, wieviele Einzelwörter aufgrund ihrer syntaktischen Funktion linguistisch gesehen als Phrasen funktionieren, in NEGRA aber keine explizite NP-Konstituente annotiert erhalten.

Einzelwort-NP in Satzgliedfunktion Um unäre NP in der Funktion von Satzgliedern und Nominalattributen miteinzubeziehen, werden in NEGRA alle Token mit den möglichen nominalen STTS-Wortarten anhand ihres Funktionstags herausgefiltert. Einbezogen sind die folgenden Satzglieder[7]: Subjekt (SB), Akkusativobjekt (OA), Dativobjekt (DA), Prädikativergänzung (PD), Genitivobjekt (OG),

[7]Die Funktionen Platzhalter (PH) und wiederholtes Element bei Platzhalterkonstruktionen (RE) führen wie in Beispiel (i) illustriert zusammen immer zu einer phrasalen Konstituente und werden in dieser Auswertung nicht noch einzeln einbezogen.

(i)



[N_{455}]

Akkusativobjekt 2 (OA2). Bei den Nominalattributen: pränominaler Genitiv (GL), postnominaler Genitiv (GR).

Die Tabelle 2.3 auf der nächsten Seite gibt einen Überblick, welche Wortarten mit welchen Funktionen in den total 15313 Fällen gepaart sind. Aus Platzgründen sind in der Tabelle alle Kombinationen mit weniger als 4 Vorkommen nicht gezeigt – es fliesssen jedoch alle 15313 Fälle in die prozentualen Auswertungen mit ein.

Pränominale und postnominale Einzelwort-Genitive beschränken sich fast ausschliesslich auf Eigennamen (NE). Bei den GL sind nebst falsch annotierter Eigennamen, ein Werktitel (10a) und Erwähnungen von Objekten mit starker emotionaler Bindung (10b) oder in Kombination davon (10c) zu finden.

(10) a. [...] so daß man an das Schicksal Wendlas aus Wedekinds" [$_{NN-GL}$ Frühlings] Erwachen" denkt. [N$_{8642}$]

b. Wenn [$_{NN-GL}$ Babys] Herz abrupt aussetzt [N$_{12015}$]

c. "Deiner Hände und unserer Hände Werk ([$_{NN-GL}$ Gottes] Schöpfung und die menschliche Verantwortung) " lautet das Thema [...] [N$_{16091}$]

Bei den GR gibt es in juristischen Verweisen eine Tendenz die Gesetzesquellen-Akronyme wie in (11) auch ohne expliziten Artikel als Genitiv zu „lesen“.

(11) So ermöglicht beispielsweise [$_{NP}$ der Paragraph 129 [$_{NN-GR}$ StGB] (kriminelle Vereinigung)] selbst bei Diebstählen die Telefonüberwachung. [N$_{2781}$]

Die Auflistung (12) gibt zusätzlich noch die Verteilung der STTS-Kategorien auf diese 15313 Fälle an. Normale Substantive weisen darin ein erstaunlich hohes Aufkommen auf.

(12) Absolute und relative Häufigkeit der STTS-Kategorien bei den impliziten Einzelwort-NP:
PPER (4965, 32.4%), NN (2203, 14.4%), PRF (2193, 14.3%), NE (1933, 12.6%), PRELS (1913, 12.5%), PIS (861, 5.6%), PDS (836, 5.5%), PWS (378, 2.5%), CARD (25, 0.2%), XY (4, 0.0%), PPOSS (2, 0.0%)

Die Tabelle 2.4 auf Seite 19 zeigt die damit bereinigten Verhältnisse zwischen koordinierten und nicht-koordinierten Nominalphrasen an. Der Anteil der CNP sinkt damit von knapp 11% (vgl. 5 auf Seite 12) auf gut 8%.

2.2.1.2 In PP eingebettete Nominalphrasen

In diesem Abschnitt soll geklärt werden, wieviele Nominalphrasen in NEGRA implizit in PP einkodiert sind. Da in den allermeisten PP eine Nominalphrase steckt, könnte man als schnelle Lösung alle gut 35500 PP (vgl. die NEGRA-Tabelle 2.2 auf Seite 14) auch als NP zählen. Dagegen sprechen folgende Gründe:

1. Eine CNP, welche als einzige nominale Konstituente in einer PP steckt, würde damit als NP gezählt. In der Tabelle 2.5 auf Seite 19 sind die 1610 Fälle zusammengestellt, welche im NEGRA-Korpus darunter fallen und knapp 5% ausmachen.

Funktion	Anzahl	in %	Wortart	Anzahl	in %
SB	9783	63.9	PPER	4235	27.7
			PRELS	1600	10.4
			NE	1316	8.6
			NN	1049	6.9
			PIS	660	4.3
			PDS	616	4.0
			PWS	284	1.9
			CARD	19	0.1
			XY	4	0.0
OA	3972	25.9	PRF	1896	12.4
			NN	879	5.7
			PPER	413	2.7
			PRELS	271	1.8
			PDS	199	1.3
			PIS	156	1.0
			PWS	84	0.5
			NE	69	0.5
DA	773	5.0	PPER	303	2.0
			PRF	297	1.9
			NE	46	0.3
			NN	43	0.3
			PRELS	38	0.2
			PIS	27	0.2
			PDS	16	0.1
GL	288	1.9	NE	280	1.8
			NN	6	0.0
PD	265	1.7	NN	217	1.4
			PPER	13	0.1
			PIS	11	0.1
			NE	10	0.1
			PWS	6	0.0
GR	226	1.5	NE	212	1.4
			PIS	7	0.0
			NN	7	0.0
OG	3	0.0	PRELS	1	0.0
OA2	3	0.0	PWS	1	0.0

Tabelle 2.3: Verteilung der 15313 impliziten NP mit Satzgliedfunktion in NEGRA. Die Anzeige der Zeilen ist auf Mindestvorkommen von 4 in Spalte 5 eingeschränkt, aber alle Fälle sind in die prozentualen Anteile eingerechnet.

Funktion	Anzahl	in %	Typ	Anzahl	in %
C(NP)	62768	100.0	NP	42286	67.4
			N	15313	24.4
			CNP	5169	8.2

Tabelle 2.4: Verhältnis der impliziten Einzelwort-Nominalphrasen (N) zu den annotierten NP und CNP in NEGRA

in %	Anzahl	Tochterkonstituenten	kumulativ
92.0	1481	APPR-AC CNP-NK	92
4.1	66	APPRART-AC CNP-NK	96
2.1	34	ADV-MO APPR-AC CNP-NK	98
0.5	8	APPR-AC APPR-AC CNP-NK	99
0.4	6	KOKOM-CM APPR-AC CNP-NK	99

Tabelle 2.5: Verteilung der 1610 PP mit nicht-erweiterten CNP in NEGRA. Angezeigt werden Tochterkonstituenten mit 3 Mindestvorkommen.

2. Es gibt PP, welche keine morphologisch separaten nominale Teile enthalten. Diese haben als Kern ein Pronominaladverb (PROAV)[8], welches sich innerhalb von PP meistens mit einem resumptiven Element (RE) verbindet, welches ein finiter oder infiniter Satz ist wie in (13).

   In der Tabelle 2.6 auf der nächsten Seite findet sich eine Zusammenstellung der häufigsten aller 383 relevanten Fälle, welche insgesamt nur 1% ausmachen. Auffällig, weil keine resumptive Konstruktion, sind die Konstruktionen, welche mit PROAV-AC annotiert sind. Es geht dabei um Formulierungen wie „darüber hinaus“, welche eine Zirkumposition ausdrücken im Sinn von „über es hinaus“. Die Annotation betont mit dem Funktionslabel AC den Präpositionalencharakter des Pronominaladverbs (oder eben Präpositionaladverbs).

(13) Stolz sind die drei Dirigenten vor allem [$_{PP}$ [$_{PROAV-PH}$ darauf] , [$_{S-RE}$ daß sie ihr Konzert ohne „eingekaufte“ Musiker bestreiten]]. [N_{2940}]

Insgesamt finden sich so noch 33528 implizit in PP eingebettete nicht-koordinierte NP. Der Bau der PP mit eingebetteten NP ist extrem vielfältig. Wenn die funktionale Bestimmung miteinbezogen wird, gibt es in NEGRA 2082 verschiedene PP-Typen mit einem Type-Token-Verhältnis, das 1:16.1 beträgt. Die Tabelle 2.7 zeigt die Verteilung derjenigen Phrasen, welche mindestens 100 Vorkommen aufweisen in NEGRA.

Der Anteil der koordinierten NP würde also mit dem Einbezug der 33528 in PP eingebetteten nur noch 5169 Fälle von total 96296 ausmachen, d.h. knapp 5.4%.

[8]Der Grammatik-Duden (Dudenredaktion 2005, §858) verwendet für die Wortart den Begriff „Präpositionaladverb“ und reserviert den Ausdruck „Pronominaladverb“ für die Satzgliedfunktion.

in %	Anzahl	Tochterkonstituenten	kumulativ
46.5	178	PROAV-PH S-RE	46
26.6	102	PROAV-PH VP-RE	73
4.2	16	PROAV-AC APZR-AC	77
3.9	15	PROAV-PH CS-RE	81
2.9	11	ADV-MO PROAV-PH S-RE	84
2.3	9	S-RE PROAV-PH	86
2.3	9	PROAV-PH CVP-RE	89
2.3	9	ADV-MO PROAV-PH VP-RE	91
1.0	4	PP-MO PROAV-PH VP-RE	92
1.0	4	PP-MO PROAV-PH S-RE	93
1.0	4	ADV-MO ADV-MO PROAV-PH S-RE	94

Tabelle 2.6: Alle 383 PP ohne Tochter mit NK-Funktion in NEGRA

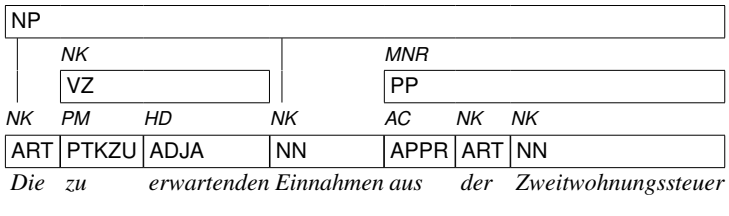
2.2.1.3 Implizite Einzelwort-AP

In NP eingebettete attributive Adjektive Das Annotationshandbuch von NEGRA (Brants u. a. 1999, 7ff.) beschreibt eine NP als bestehend „aus einer Reihe von pronominalen, substantivischen und adjektivischen Kernelementen“. Attributive Adjektive sind deshalb grundsätzlich flach eingebettet in NP und auch in PP, ausser sie haben Ergänzungen oder werden selbst modifiziert. Alle attributiven Adjektive, welche nicht schon Kopf (HD) einer AP sind, können implizit als AP-erzeugend betrachtet werden. Damit ergeben sich zusätzlich 17725 AP-Kandidaten, deren Mutterkategorien in der Auflistung (14) ausgewiesen sind.

(14) Verteilung der ADJA in NEGRA ohne HD-Funktion:
NP (10680, 60.3%), PP (6925, 39.1%), VZ (29, 0.2%), MPN (27, 0.2%), MTA (24, 0.1%), NM (12, 0.1%), S (7, 0.0%), CNP (7, 0.0%), AA (7, 0.0%), VP (3, 0.0%), AVP (2, 0.0%), ISU (1, 0.0%), CO (1, 0.0%)

Anzahlmässig beinahe vernachlässigbar sind alle Fälle, bei denen die Mutterkategorie weder NP noch PP ist. Die Fälle mit der Mutterkategorie VZ betreffen departizipiale Adjektiv-Konstruktionen wie in (15a). Die Mehrwortlexem-Kategorie MPN ist fast durchwegs dem Token „St.“ wie in (15b) zuzuschreiben. Die Mehrwortlexem-Kategorie MTA tritt bei Herkunftsadjektiven auf, welche aus mehrteiligen Ortsbezeichnungen gebildet sind wie in (15c).

(15) a.



[N_{13194}]

b. Goodwill Games 1994 in [$_{MPN}$ [$_{ADJA-PNC}$ St.] Petersburg] [N_{14670}]

in %	Anzahl	Tochterkonstituenten	kum.
11.4	3822	APPR-AC ART-NK NN-NK	11
8.2	2733	APPRART-AC NN-NK	20
8.1	2703	APPR-AC NN-NK	28
6.2	2078	APPR-AC NE-NK	34
4.5	1519	APPR-AC ART-NK ADJA-NK NN-NK	38
3.9	1310	APPR-AC ADJA-NK NN-NK	42
3.0	1017	APPR-AC CARD-NK NN-NK	45
3.0	1007	APPR-AC ART-NK NN-NK NP-GR	48
2.2	742	APPRART-AC ADJA-NK NN-NK	50
2.2	730	APPR-AC ART-NK NN-NK PP-MNR	53
1.6	546	APPRART-AC NN-NK NP-GR	54
1.4	471	APPR-AC PPOSAT-NK NN-NK	56
1.3	432	APPR-AC NN-NK NP-GR	57
1.2	404	APPR-AC NN-NK PP-MNR	58
1.1	385	APPR-AC AP-NK NN-NK	59
1.1	378	APPR-AC CARD-NK	60
1.1	374	APPR-AC PDAT-NK NN-NK	62
1.0	352	APPRART-AC NN-NK PP-MNR	62
1.0	348	APPR-AC PRELS-NK	64
1.0	335	APPR-AC MPN-NK	64
0.9	289	APPR-AC ART-NK NE-NK	65
0.8	273	APPR-AC PIS-NK	66
0.7	231	APPR-AC ART-NK ADJA-NK NN-NK PP-MNR	67
0.6	214	APPR-AC ART-NK ADJA-NK NN-NK NP-GR	68
0.6	198	APPR-AC PPER-NK	68
0.6	190	APPRART-AC NE-NK	69
0.5	183	APPR-AC ART-NK AP-NK NN-NK	69
0.5	171	APPR-AC PIDAT-NK NN-NK	70
0.5	162	APPR-AC NM-NK NN-NK	70
0.5	159	APPR-AC ART-NK NN-NK S-RC	71
0.4	150	APPR-AC ART-NK NN-NK PP-PG	71
0.4	139	APPR-AC ADJA-NK NN-NK PP-MNR	72
0.4	139	ADV-MO APPR-AC NN-NK	72
0.4	136	ADV-MO APPR-AC ART-NK NN-NK	72
0.4	132	APPR-AC CAP-NK NN-NK	73
0.4	120	APPR-AC PPOSAT-NK ADJA-NK NN-NK	73
0.3	116	APPRART-AC NN-NK CARD-NK	73
0.3	115	APPR-AC ADV-NK	74
0.3	111	APPR-AC ART-NK NN-NK NE-NK	74
0.3	105	APPRART-AC NN-NK NP-APP	74
0.3	104	APPR-AC PIAT-NK NN-NK	75

Tabelle 2.7: Verteilung der 33528 PP mit implizit eingebetteten NP in NEGRA. Ausgeschlossen sind die Fälle, welche als nominales Element nur CNP enthalten. Gezeigt werden nur Tochterkonstituenten mit mindestens 100 Vorkommen.

c. Mit einem einzigen Fahrschein von Grävenwiesbach nach Darmstadt oder vom [$_{MTA}$ [$_{NE-MTA}$ Bad] [$_{ADJA-MTA}$ Homburger]] Kurhaus nach Wiesbaden : [N$_{17486}$]

Mehrfachvorkommen von Adjektiven und adjektivischen Konstituenten innerhalb von annotierten Phrasen In der Auflistung (14) auf Seite 20 wurden attributive Adjektive, welche in derselben Phrase vorkommen, jeweils einzeln gezählt. In diesem Abschnitt sollen die Konstruktionen mit mehr als einem Adjektiv genauer betrachtet werden.

Bei den restriktiven Adjektiven gibt es bekanntlich zwei semantisch unterschiedliche Konstruktionen, welche auch in der Interpunktion bzw. Pausensetzung differenziert sind. So modifiziert und restringiert in (16a) das Adjektiv „religiöser" das Kernnomen „Zusammenstöße". Diese Gruppe wiederum wird als Ganzes modifiziert von „schwerer". Die beiden Adjektive bilden somit keine Einheit zusammen, sondern stehen im Modifikationsverhältnis. Im Gegensatz dazu steht die Konstruktion in (16b), welche der Grammatik-Duden (Dudenredaktion 2005, §1285) als Reihung und (Zifonun u. a. 1997, 1992) explizit als kooordinativ mit der Bedeutung „und" bestimmt. Hier bilden die beiden Adjektive eine Einheit, welche durch das Komma bzw. die Pause gesetzt wird. In unserer Terminologie liegt bei (16b) klar eine asyndetische Adjektiv-Koordination vor.

(16) a. In Zeiten der Reformation und [$_{NP}$ [$_{ADJA-NK}$ schwerer] [$_{ADJA-NK}$ religiöser] [$_{NN}$ Zusammenstöße]] [. . .] [N$_{153}$]

b. Es wird [$_{PP}$ [$_{APPR-AC}$ in] [$_{ADJA-NK}$ kleinen] , [$_{ADJA-NK}$ umweltfreundlichen] [$_{NN}$ Betrieben]] gearbeitet. [N$_{993}$]

c. [. . .] Fortschritte im Bereich Kinderbetreuung [$_{PP}$ durch das [$_{ADJA-NK}$ neue] , [$_{ADJA-NK}$ dreigeteilte] [$_{ADJA-NK}$ städtische] Programm] .[N$_{398}$]

Im NEGRA-Korpus sind solche asyndetische Adjektive-Koordinationen jedoch selten als CAP annotiert, obwohl das NEGRA-Annotationshandbuch (Brants u. a. 1999, 86) explizit mahnt: „Beachte: die Präsenz einer koordinierenden Konjunktion ist nicht notwendig. Aufzählungen werden ebenso annotiert."

In der Tabellenzusammenstellung 2.8 auf der nächsten Seite sind die adjektivischen Konstituenten, welche innerhalb einer Phrase vorkommen, paarweise im Detail mit ihrer Kategorie aufgeschlüsselt. Die Spalte „±-Komma" spezifiziert dabei, ob zwischen den beiden Elementen ein Komma liegt oder nicht. Das Vorkommen von Kommas ist ein sehr starkes Indiz, dass es sich bei diesem Paar eigentlich um eine koordinierte Adjektivphrase handelt und somit ein Annotationsproblem vorliegt. Gezählt wurden immer Paare von einem attributiven Element und dem nächst nachfolgenden. In einem Satz wie (16c) gibt es einen koordinativen und einen nicht-koordinativen Fall innerhalb der PP.

Insgesamt gibt es im NEGRA-Korpus 166 Fälle, wo mehrere ADJA, AP oder CAP eigentlich koordiniert annotiert sein müssten wie im Beispiel (16b). Immerhin mehr als 1/6 der 900 explizit als CAP annotierten Strukturen. Der Fall aus Beispiel

NEGRA (total 1097)

in %	Anzahl	Konstituente 1	Konstituente 2	± Komma	kumulativ
59.7	655	ADJA	ADJA	–	60
19.6	215	AP	ADJA	–	79
7.7	84	ADJA	ADJA	+	87
3.8	42	ADJA	AP	+	91
2.3	25	ADJA	AP	–	93
2.2	24	AP	AP	+	95
1.4	15	AP	ADJA	+	97
1.2	13	ADJA	CAP	–	98
0.9	10	CAP	ADJA	–	99
0.7	8	AP	AP	–	100
0.5	5	AP	CAP	–	100
0.1	1	ADJA	CAP	+	100

TIGER (total 2194)

in %	Anzahl	Konstituente 1	Konstituente 2	± Komma	kumulativ
69.1	1517	ADJA	ADJA	–	69
21.7	476	AP	ADJA	–	91
1.9	42	AP	AP	–	93
1.8	40	ADJA	CAP	–	94
1.8	39	ADJA	AP	–	96
1.6	35	CAP	ADJA	–	98
0.9	19	ADJA	AP	+	99
0.7	15	AP	CAP	–	100
0.1	3	AP	AP	+	100
0.1	3	AP	ADJA	+	100
0.1	2	CAP	AP	–	100
0.1	2	ADJA	ADJA	+	100
0.0	1	CAP	CAP	–	100

Tabelle 2.8: Verteilung der adjazenten annotierten Adjektiv-Konstituenten mit bzw. ohne Komma dazwischen in NEGRA und TIGER.
Lesebeispiele für die 3. Zeilen: In NEGRA gibt es 84 Fälle, wo innerhalb einer Phrase zwei attributive Adjektive (ADJA) nur durch ein Komma getrennt nebeneinander vorkommen. In TIGER gibt es 42 Fälle, wo innerhalb einer Phrase 2 Adjektivphrasen (AP) ohne Komma dazwischen unmittelbar nebeneinander vorkommen.

(16a), wo die Adjektive in einer Modifikationsbeziehung zueinander stehen, hat insgesamt 931 Vorkommen in NEGRA.

Auffällig ist weiter, dass bei den nicht-koordinativen Verknüpfungen komplexe Phrasen (AP) häufiger als erstes Element vorkommen denn als zweites. Bei den koordinativen Adjektiven ist das „schwerere“ Element im Gegensatz dazu gerade häufiger an zweiter Stelle zu finden.

Eine Auswertung über dem TIGER-Korpus in derselben Tabellenzusammenstellung 2.8 auf der vorherigen Seite bestätigt diese Tendenz. Immerhin halten sich die Annotatoren dort deutlich konsequenter an die im TIGER-Annotationsschema (Albert u. a. 2003, 112) ebenfalls vermerkte Mahnung, dass Reihungen als Koordination zu annotieren sind. Nur noch 27 Fälle gibt es, wo 2 nur durch Komma getrennte Adjektiv-Konstituenten nicht koordiniert annotiert werden.

Ganz im Gegensatz dazu stehen die Annotationen multipler Adjektive im CZ-Korpus. Dort gibt es in den 3000 Sätzen nur 4 multiple Adjektive, dafür 238 CAP. Die Annotationsstrategie ist hier gerade umgekehrt zu TIGER und NEGRA, denn hier werden auch alle einander modifizierenden Adjektive als Koordination annotiert.

2.2.2 Koordinationsmittel

Eine umfassende deskriptive Studie zu den Konjunktoren liegt in der deutschen Sprachwissenschaft mit Pasch (2003) vor. Einiges an Material zu den Konjunktoren davon ist im Online-Grammatik-Portal „GRAMMIS“ des Instituts für deutsche Sprache (IDS 2006) zugänglich. Im Folgenden werden die wichtigsten Eigenschaften erhoben, welche sich mittels annotierter Korpora quantifizieren lassen.

Wie werden die Konjunkte in koordinierten Konstruktionen verknüpft? Die wichtigste Unterscheidung ist diejenige zwischen lexikalisierten Koordinationsmitteln, welche sich vorwiegend aus nebenordnenden Konjunktoren rekrutieren und als Wortart das STTS-Kürzel KON tragen, und Koordinationsmitteln, welche nur aus Interpunktion bestehen und als „Wortart“ das STTS-Tag $, (nur für Kommata) bzw. $(für weitere satzinterne Interpunktionen wie Gedankenstrich aufweisen. Gemäss dem NEGRA-Annotationsmodell, das Interpunktion nie in die syntaktische Struktur einbaut, finden sich nur die lexikalisierten Konjunktoren in die Satzstruktur integriert.

Syndetische Koordination liegt vor, wenn nur lexikalische Koordinationsmittel eingesetzt werden, während asyndetische Koordination vorliegt, wenn nur Interpunktion verwendet wird. Eine Mischform, welche syndetische und asyndetische Koordinationsmittel einsetzt, stellt die monosyndetische Koordination dar, wobei der lexikalische Konjunktor immer die beiden letzten Konjunkte verknüpfen muss.

Die Tabelle 2.9 auf der nächsten Seite zeigt, wie sich die 3 verschiedenen Typen in NEGRA und TIGER insgesamt verteilen. Die syndetischen Koordinationen machen etwas mehr als 7/10 aus, die monosyndetischen etwa 1/10, die asyndetischen etwas weniger als 2/10.

Die Koordinationstypen mit dem Kürzel „x“ entstehen durch unterschiedliche

NEGRA			TIGER		
in %	Anzahl	Typ	in %	Anzahl	Typ
72.1	7163	syn	72.8	14083	syn
16.8	1666	asyn	17.0	3299	asyn
10.5	1043	mono	9.7	1872	mono
0.7	69	x	0.5	97	x

Tabelle 2.9: Verteilung der syndetischen, asyndetischen und monosyndetischen Koordination in NEGRA und TIGER. Legende zu den Koordinationstypen: syn=syndetisch; asyn=asyndetisch; mono=monosyndetisch; x=andere

Mischformen wie im orthographisch problematischen Beispiel (17a) bzw. (17b) auf Seite 25, das ebenfalls paarige syndetische Koordination von Teil-Konjunkten zeigt, oder durch Annotationsfehler (insbesondere die Verwechslung des Funktionskürzel CD für Konjunktoren mit CJ für Konjunkt). Oder es sind Artefakte der Auswertung in diskontinuierlichen Strukturen: So entsteht in Beispiel (17c) durch den diskontinuierlichen Einschub des 1. Teils des paarigen Konjunktors „entweder“ eine spezielle Konfiguration.

(17) a.

CNP								1	
CJ		*CD*	*CJ*		*CJ*	*CD*	*CJ*		
NN	$,	KON	NN	$,	NN	KON	NN	VMFIN	0 [. . .] [N_{2030}]
Mehl ,		*und*	*Zucker ,*		*Milch und*		*Butter müssen*		

b. Zusammen mit rund 150 anderen Mädchen lernt sie dort [$_{CNP}$ Buchhaltung und Kochen, Schreibmaschineschreiben, Weißnäherei und die Stickerei] . [N_{3225}]

c. [. . .] indem man diese Fremden [. . .] darstellt und [$_{CS}$ [$_{S}$ sie ⌊ [$_{KON-CD}$ entweder] ⌋ beschwört] [$_{KON-CD}$ oder] [$_{S}$ auf ganz eigene Weise interpretiert]] , [. . .] [N_{206}]

Die detailliert Verteilung dieser Koordinationstypen über den verschiedenen koordinierten Phrasen in NEGRA zeigt die Tabelle 2.10 auf der nächsten Seite. Die syndetischen Koordinationen sind entsprechend ihres absoluten Übergewichts von 70% auch bezüglich der einzelnen Phrasen die häufigste Konstruktion. Bei den CNP sind monosyndetische Koordinationen gegenüber asyndetischen klar bevorzugt, während bei den CS, CVP, CPP sowie CO der Fall gerade umgekehrt liegt. Bei CAP und CAVP sind die Verhältnisse ausgeglichen.

Dieselbe Auswertung für TIGER in Tabelle 2.11 auf Seite 27 zeigt ein ähnliches Bild, ausser bei CAP und CAVP, wo asyndetische Koordination die monosyndetische eindeutig dominiert. Bei den CAP liegen die Gründe in der unterschiedlichen Annotation von Adjektiven, welche mit Kommas gereiht sind und im Abschnitt 2.2.1.3 auf Seite 20 bereits analysiert wurden.

in %	Anzahl	in %	Koordinationstyp	Anzahl	in %
CNP	5169	52.0	syn	3959	39.8
			mono	742	7.5
			asyn	422	4.2
			x	46	0.5
CS	2555	25.7	syn	1408	14.2
			asyn	978	9.8
			mono	152	1.5
			x	17	0.2
CAP	900	9.1	syn	769	7.7
			mono	66	0.7
			asyn	63	0.6
			x	2	0.0
CVP	547	5.5	syn	441	4.4
			asyn	70	0.7
			mono	33	0.3
			x	3	0.0
CPP	477	4.8	syn	345	3.5
			asyn	94	0.9
			mono	37	0.4
			x	1	0.0
CO	166	1.7	syn	122	1.2
			asyn	35	0.4
			mono	9	0.1
CAVP	95	1.0	syn	87	0.9
			mono	4	0.0
			asyn	4	0.0
CAC	26	0.3	syn	26	0.3
CVZ	5	0.1	syn	5	0.1
CCP	1	0.0	syn	1	0.0

Tabelle 2.10: Verteilung der Koordinationstypen in NEGRA. Legende zu den Koordinationstypen: syn=syndetisch; asyn=asyndetisch; mono=monosyndetisch; x=andere

in %	Anzahl	in %	Koordinationstyp	Anzahl	in %
CNP	9893	51.1	syn	7711	39.8
			mono	1366	7.1
			asyn	772	4.0
			x	44	0.2
CS	4702	24.3	syn	2774	14.3
			asyn	1631	8.4
			mono	266	1.4
			x	31	0.2
CAP	1894	9.8	syn	1318	6.8
			asyn	479	2.5
			mono	94	0.5
			x	3	0.0
CVP	1283	6.6	syn	1064	5.5
			asyn	124	0.6
			mono	82	0.4
			x	13	0.1
CPP	1051	5.4	syn	786	4.1
			asyn	213	1.1
			mono	48	0.2
			x	4	0.0
CO	309	1.6	syn	226	1.2
			asyn	67	0.3
			mono	13	0.1
			x	3	0.0
CAVP	177	0.9	syn	165	0.9
			asyn	11	0.1
			mono	1	0.0
CAC	21	0.1	syn	18	0.1
			asyn	2	0.0
			mono	1	0.0
CVZ	20	0.1	syn	19	0.1
			mono	1	0.0
CCP	1	0.0	syn	1	0.0

Tabelle 2.11: Verteilung der Koordinationstypen in TIGER. Legende zu den Koordinationstypen: syn=syndetisch; asyn=asyndetisch; mono=monosyndetisch; x=andere

syndetisch			asyndetisch			monosyndetisch		
in %	Anzahl	CD	in %	Anzahl	CD	in %	Anzahl	CD
99.3	22427	2	82.7	4351	2	75.0	2411	3
0.6	146	3	12.7	670	3	17.0	545	4
0.0	7	4	3.1	164	4	4.8	154	5
0.0	1	5	0.8	42	5	1.6	52	6
			0.3	17	6	0.7	22	7
			0.1	6	8	0.5	15	8
			0.1	4	7	0.3	9	9
			0.1	3	10	0.1	3	11
			0.0	2	13	0.1	2	10
			0.0	1	9	0.0	1	13

Tabelle 2.12: Verteilung der Anzahl Konjunkte (CD) in NEGRA, TIGER und CZ aufgeschlüsselt nach Koordinationstyp. Mittelwerte mit Standardabweichung: syndetisch: 2.01 $\pm$ 0.09 , asyndetisch: 2.25 $\pm$ 0.68 , monosyndetisch: 3.4 $\pm$ 0.89

2.2.2.1 Konjunktanzahl

Mit wie vielen Konjunkten muss in den verschiedenen Koordinationstypen gerechnet werden? Da monsyndetische Koordinationen per Definition mindestens 3 Konjunkte enthalten, macht eine Aufschlüsselung bezüglich der Koordinationstypen Sinn. Für die folgende Auswertung wurden alle Daten aus NEGRA, TIGER und CZ zusammengenommen, was bei insgesamt 31251 Vorkommen die folgende Verteilungen für die Koordinationstypen ergibt: syndetisch (22581, 72.3%), asyndetisch (5260, 16.8%), monosyndetisch (3214, 10.3%), andere (196, 0.6%) .

Die Tabellenzusammenstellung 2.12 zeigt, dass syndetische Koordinationen zu über 99% genau 2 Konjunkte enthalten und äusserst homogen sind. Die asyndetischen Koordinationen mit 2 Konjunkten decken gut 82% der Fälle ab, 3 Konjunkte bilden mit knapp 13% eine ansehnlich Nebengruppe. Diese beiden häufigsten Klassen ergeben kumulativ schon über 95% aller Fälle. Bei einer Standardabweichung von 0.68 vom Mittelwert 2.25 sind sie allerdings noch homogener als die kleinste Gruppe der monosyndetischen Koordinationen. Exakt 3 Konjunkte haben dabei „nur“ noch 75% und die grössere Variabilität wird auch durch die Standardabweichung von 0.89 bei einem Mittelwert von 3.4 belegt. Aber auch hier decken die beiden häufigsten Klassen mit 92% den Grossteil der Fälle ab.

2.2.2.2 Konjunktoren

Welche Konjunktoren finden sich in den untersuchten Korpora wie oft? Bei der Antwort auf diese Frage muss berücksichtigt werden, dass eine Koordination mehr als einen Konjunktor enthalten kann und einige wenige Konjunktoren paarig als

Klammerstruktur funktionieren.

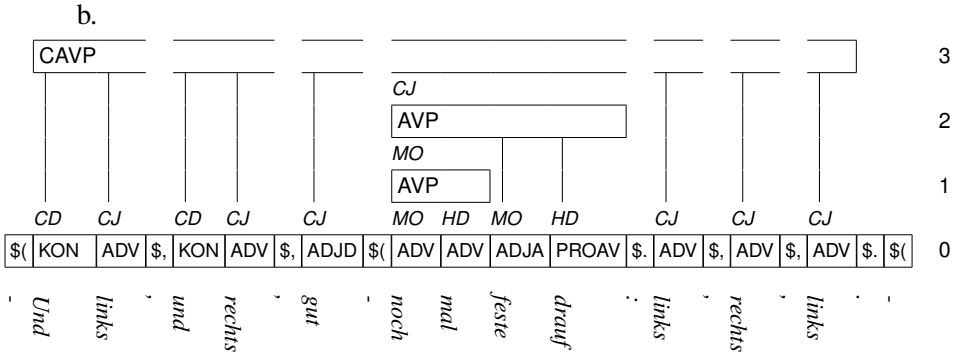
Paarige Konjunktoren In Pasch (2003, 471ff.) werden die komplexen Distributionseinschränkungen paariger Konjunktoren detailliert diskutiert. Da diese Fälle marginal sind, werden sie hier separat abgehandelt. Während bei den Präposition die paarigen Zirkumpositionen in STTS für das Vorder- und Nachglied (APPR bzw. APPZR) zwei verschiedene Tags erhalten, wird eine solche Unterscheidung nicht gemacht. Paarige Konjunktoren muss man deshalb konfigurational erschliessen, indem man prüft, ob in koordinierten Phrasen die 1. Konstituente in CD-Funktion vor der 1. Konstituente in CJ-Funktion steht. Das Problem dabei sind die Fälle wie in (17c) auf Seite 25, wo der einleitende Konjunktor nicht vor dem 1. Konjunkt, sondern diskontinuierlich zwischendrin steht. In (18) sind die total 85 paarig annotierten Konjunktorsequenzen aufgelistet, wobei in 73 Fällen der einleitende Konjunktor diskontinuierlich im 1. Konjunkt erscheint.

(18) Paarig annotierte Konjunktoren in NEGRA:
„weder noch“ (30, 35.3%), „sowohl als“ (26, 30.6%), „entweder oder“ (14, 16.5%), „sowohl wie“ (8, 9.4%), „sowohl und“ (2, 2.4%), „weder noch oder“ (1, 1.2%), „weder noch noch“ (1, 1.2%), „und und“ (1, 1.2%), „ob oder“ (1, 1.2%), „aber sondern“ (1, 1.2%)

Neben den 4 erwarteten und auch häufigsten Standardfällen (Brants u. a. 1999, 86) tauchen noch Spezialfälle auf. Satz (19a) enthält eine Variante von „sowohl als/wie (auch)“[9], die im Grammatischen Wörterbuch (IDS 2006) nicht aufgeführt ist. Ein anderes Problem sind initiale Konjunktoren, welche den Fällen mit „und und“ sowie „aber sondern“ zugrunde liegen und normalerweise mit der grammatischen Funktion Junktor JU versehen werden. In (19b) ist der eine Fall aufgeführt, der einer speziellen Äusserung entstammt und grundsätzliche Fragen aufwirft, wie solche Sprechakte kodiert werden sollen.

(19) a. Ganz am Rande, **sowohl** örtlich **und auch** wertungsmäßig gesehen , [. . .]
[N_{5676}]

b.



[9]Das „auch“ in „sowohl als/wie/und auch“ wird im NEGRA-Annotationsschema nie als Konjunktorbestandteil aufgefasst.

[N_{5314}]

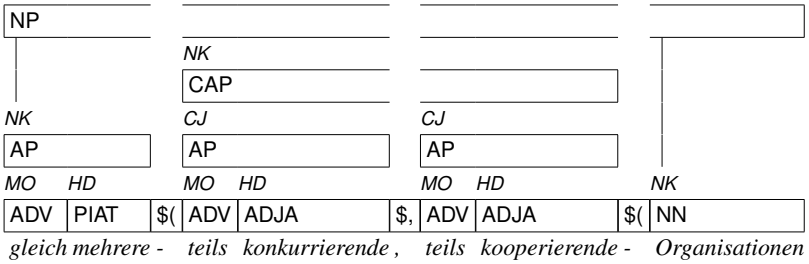
In TIGER zeigt sich bei total 201 Fällen ein ähnliches Bild, allerdings ist dort „weder noch“ deutlich stärker vertreten. Bei 177 Fällen erscheint dabei der einleitende Konjunktor diskontinuierlich im 1. Konjunkt.

(20) Paarig annotierte Konjunktoren in TIGER:
„weder noch“ (96, 47.8%), „sowohl als“ (56, 27.9%), „sowohl wie“ (21, 10.4%), „entweder oder“ (15, 7.5%), „sowohl als und“ (3, 1.5%), „sowohl wie und“ (2, 1.0%), „zwar aber“ (1, 0.5%), „wenn dann“ (1, 0.5%), „weder noch sondern“ (1, 0.5%), „weder noch oder“ (1, 0.5%), „weder noch noch“ (1, 0.5%), „sowohl —“ (1, 0.5%), „aber und“ (1, 0.5%), „— sondern auch“ (1, 0.5%)

Auch in TIGER werden die sogenannt paarig klammernden Konjunktoren in seltenen Fällen erweitert zu ternären Strukturen.

Paarige Adverbkonnektoren Neben den Konjunktoren existieren weitere meist paarig auftretende Formen, welche in Pasch (2003) „Adverbkonnektoren“ genannt werden und in NEGRA als modifizierende Adverbien in die Konjunkte eingebaut annotiert werden wie in (21).

(21) Aus den Kämpfen gegen Großgrundbesitzer oder Bodenschatz-Sucher haben sich in Kolumbien



entwickelt. [N_{1886}]

Typisch ist dabei die reduplizierende Füllung: „bald... bald“, „halb... halb“, „mal... mal“, „sei es... sei es“, „teils... teils“. Die untersuchten Korpora enthalten jedoch keine oder nur Einzelvorkommen davon.

Mehrwort-Konjunktoren Im Gegensatz zu den paarig klammernden Konjunktoren setzen sich Mehrwort-Konjunktoren immer aus adjazenten Wörtern zusammen, welche als Gruppe in Konjunktorfunktion CD stehen. Sie sind in NEGRA mit 2 Fällen sehr spärlich annotiert. Der eine Fall ist „geschweige denn“ in (22a). Dieser Konjunktor kommt nochmals vor in NEGRA, wird dort aber als Einzelkonjunktor behandelt. Der andere Mehrwort-Konjunktor ist „wohl aber“, der sich in (22b) schon fast paarig mit „zwar“ verbindet als Adversativ-Konstruktion.

(22) a. Die stehen den Männer Hekmatyars näher als den tadschikischen Partisanen Massuds, [$_{AVP-CD}$ [$_{KON-AVC}$ geschweige] [$_{ADV-AVC}$ denn]] den usbekischen Milizionären. [N$_{4172}$]

b. Was wie der Turm einer alten Wehrkirche aussieht, war zwar keine Kirche, [$_{AVP}$[$_{ADV-MO}$ wohl] [$_{KON-HD}$ aber]] eine mittelalterliche Wehranlage [...] [N$_{2073}$]

In TIGER sind mit 7 Mehrwort-Konjunktoren zwar etwas mehr Fälle annotiert, aber es ergibt sich immer noch ein recht unsystematisches Bild. Neben 3 Fällen von „geschweige denn“ wird das Adverb „auch“ in (23a) und (23c), jedoch nicht in (23b) zum Konjunktor genommen.

(23) a. Auch Jägemann und Rolf Strojec von der Hessischen Kanuschule setzen vielmehr auf ein System, das ausschließlich der Natur vorbehaltene Rückzugsräume garantiert, [$_{AVP-CD}$[$_{KON-AVC}$ und] [$_{ADV-AVC}$zwar]] auch außerhalb der bereits heute festgesetzten Schutzgebiete. [T$_{15049}$]

b. Die Gates-Company hat schon Produktion und Distribution, [$_{AVP-CD}$ [$_{KON-AVC}$ aber] [$_{ADV-AVC}$ auch]] Rechnungswesen-Funktionen nach draußen vergeben. [T$_{22320}$]

c. Grüne und Deutscher Gewerkschaftsbund (DGB) sind einig, daß die Grundsysteme der sozialen Sicherung zwar bewährt, unter den jetzt veränderten Rahmenbedingungen und Anforderungen aber “sowohl reformfähig [$_{AVP-CD}$ [$_{KON-AVC}$als] [$_{ADV-AVC}$ auch]] reformbedürftig” sind. [T$_{23298}$]

Wenn Mehrwort-Konjunktoren systematisch als solche annotiert werden sollen, braucht es klare linguistische und lexikalische Vorgaben.

2.2.2.3 Syndetische Verwendung

Die syndetisch verwendeten Konjunktoren sind in der Zusammenstellung 2.13 auf der nächsten Seite ersichtlich. Da syndetische Koordination mit 2 Konjunkten extrem dominant ist, erscheinen aufgrund des Mindestvorkommens relativ wenig Zeilen mit mehr als einem Konjunktor. Bei NEGRA gibt es 59 verschiedene Konjunktoren mit einem Type-Tokenverhältnis von 1:121.4. TIGER hat mit 64 verschiedenen Konjunktoren, aber auch deutlich mehr Koordinationen, ein weitaus höheres Type-Tokenverhältnis von 1:220.0.

Aufgrund der Textsorte, welche einiges an Sportresultaten enthält, und aufgrund der Tokenisierungsstrategie erscheint der Doppelpunkt in NEGRA (und nur in NEGRA) als Interpunktionskonjunktor, welcher als solcher sogar in die syntaktische Struktur wie in Beispiel (24a) integriert wird – im Gegensatz zur sonstigen Behandlung von Interpunktion in NEGRA. In TIGER wird dagegen für numerische Verhältnisangaben wie in (24b) eine andere Tokenisierungsstrategie gewählt, welche die eher fragwürdige Annotation solcher „Koordinationen“ auf syntaktischer Ebene überflüssig macht.

NEGRA (total 7163)				TIGER (total 14083)			
in %	Anz.	Füllung	k.	in %	Anz.	Füllung	k.
79.3	5681	und	79	80.0	11269	und	80
6.5	464	oder	86	6.4	904	oder	86
3.2	227	sondern	89	3.1	432	sondern	90
2.7	192	aber	92	2.2	310	aber	92
2.4	175	bis	94	1.8	248	sowie	94
1.7	119	sowie	96	1.4	193	bis	95
0.7	51	doch	96	0.9	127	wie	96
0.4	27	wie	97	0.8	113	denn	97
0.3	24	(sowohl) als	97	0.6	82	(weder) noch	97
0.3	23	(weder) noch	98	0.6	81	doch	98
0.3	22	&	98	0.4	55	(sowohl) als	98
0.3	20	:	98	0.3	36	beziehungsweise	98
0.3	19	und und	98	0.2	24	und und	99
0.1	10	(entweder) oder	98	0.1	21	bzw.	99
0.1	10	beziehungsweise	99	0.1	20	(sowohl) wie	99
0.1	9	für	99	0.1	19	und sowie	99
0.1	8	(sowohl) wie	99	0.1	15	+	99
0.1	8	bzw.	99	0.1	10	(entweder) oder	99
0.1	7	+	99	0.1	10	sowie und	99
0.1	7	und sowie	99	0.1	9	als	99
0.1	6	sowie und	99	0.1	8	gegen	100
0.1	5	zu	99	0.0	7	oder und	100
0.1	4	mal	99	0.0	6	—	100
0.1	4	plus	99	0.0	6	und aber	100
0.0	2	—	99	0.0	6	zu	100
0.0	2	oder oder	99	0.0	5	and	100
0.0	2	respektive	99	0.0	5	respektive	100
0.0	2	un	99	0.0	5	und oder	100
0.0	2	und doch	99	0.0	4	oder oder	100
0.0	2	und sowie und	99	0.0	4	plus	100
				0.0	3	aber und	100
				0.0	3	jedoch	100
				0.0	3	mal	100
				0.0	3	noch	100
				0.0	3	sondern und	100
				0.0	2	auch	100
				0.0	2	oder sondern	100
				0.0	2	oder sowie	100
				0.0	2	und sondern	100
				0.0	2	und und und	100

Tabelle 2.13: Verteilung der lexikalischen Füllung der Konjunktoren in syndetischen Koordinationen in NEGRA und TIGER. Gezeigt werden Konjunktoren mit einem Mindestvorkommen von 2. Der öffnende Konjunktor ist bei paarigen Konjunktoren in Klammern gesetzt.

(24) a. Die Schienenrutscher waren im Halbfinale in einem dramatischen Spiel [$_{CAP}$ [$_{CARD-CJ}$ 3] [$_{\$.-CD}$:] [$_{CARD-CJ}$ 2]] gegen " Blaugold " unterlegen. [N_{3427}]

b. Immer noch gilt als allein verlässig der Maßstab [$_{CARD}$ 1:1] , [...] [T_{18336}]

Das Pluszeichen „+“ und das Kaufmanns-Und „&“ sind weitere symbolische Konjunktoren, welche in den Korpora recht häufig als Konjunktoren annotiert werden.

Korpusabhängige Schwankungen in der Häufigkeitsreihenfolge treten insbesondere bei der Verwendung von Partikeln auf, welche in STTS kein Kürzel für nebenordnende Konjunktion (KON) erhalten. Für einfache syndetische Konjunktoren trifft dies 21 Typen in 230 Fällen, welche für NEGRA in der Auflistung (25) gezeigt werden. Darunter sind auch Annotationsfehler und Inkonsistenzen zu finden.

(25) Auflistung der annotierten Konjunktoren in NEGRA ohne KON-Wortart: „bis/APPR“ (175), „:/$.“ (18), „für/APPR“ (9), „zu/APPR“ (5), „mal/ADV“ (4), „aber/ADV“ (3), „:/XY“ (2), „an/APPR“ (1), „auf/APPR“ (1), „beziehungsweise/ADV“ (1), „bzw./KOUS“ (1), „doch/ADV“ (1), „ebenso/ADV“ (1), „gegen/APPR“ (1), „gleich/ADV“ (1), „in/APPR“ (1), „nach/APPR“ (1), „plus/ADV“ (1), „um/APPR“ (1), „weil/KOUS“ (1), „wenngleich/KOUS“ (1)

In TIGER betrifft dies 11 Typen in 101 Fällen, was ein deutlich konsistenteres Bild ergibt:

(26) Auflistung der annotierten Konjunktoren in TIGER ohne KON-Wortart: „bis/APPR“ (75), „gegen/APPR“ (8), „zu/APPR“ (6), „aber/ADV“ (3), „bis/ADV“ (3), „-/APPR“ (1), „also/ADV“ (1), „beziehungsweise/ADV“ (1), „draußen/ADV“ (1), „statt/APPR“ (1), „von/APPR“ (1)

Die verschiedenen Verwendungsweisen des Wortes „bis“ sind bei beiden Korpora wichtig. In NEGRA finden sich mehr Mass- und insbesondere Zeitangaben in Veranstaltungskalendern, welche wie in (27a) oder (27b) strukturiert sind.

(27) a. Nach Angaben des US-Rechnungshofs würde es allein [$_{CAP}$ [$_{CARD-CJ}$ 300] [$_{APPR-CD}$ bis] [$_{CARD-CJ}$ 400]] Milliarden Dollar kosten, [...] [N_{742}]

b. [$_{CAP}$ [$_{CARD-CJ}$ 19.] [$_{APPR-CD}$ bis] [$_{CARD-CJ}$ 21.]] Juni 1992: Jahrestagung der Vereinigung für ökologische Wirtschaftsförderung in Wuppertal. [N_{1415}]

Die Verteilung der Wortart aller koordinierenden „bis“ unterscheidet sich zwischen NEGRA und TIGER deutlich: NEGRA: bis/APPR (175) ; TIGER: bis/KON (115), bis/APPR (75), bis/ADV (3) . Man kann sich zurecht fragen, warum Lexeme, welche als Konjunktor verwendet bzw. annotiert werden, nicht konsequenterweise mit

NEGRA

in %	Anzahl	Füllung	kum.
86.7	904	und	87
8.3	87	oder	95
3.6	38	sowie	99
0.8	8	aber	99
0.2	2	doch	100

TIGER

in %	Anzahl	Füllung	kum.
85.4	1599	und	85
8.2	153	oder	94
4.8	90	sowie	98
0.9	16	aber	99
0.2	4	doch	100
0.2	3	sondern	100
0.1	2	denn	100

CZ

in %	Anzahl	Füllung	kum.
84.7	254	und	85
8.3	25	oder	93
6.0	18	sowie	99

NEGRA, TIGER, CZ

in %	Anzahl	Füllung	kum.
85.8	2757	und	86
8.2	265	oder	94
4.5	146	sowie	98
0.8	25	aber	99
0.2	6	doch	100
0.1	4	sondern	100
0.1	2	&	100
0.1	2	denn	100
0.1	2	noch	100

Tabelle 2.14: Verteilung der lexikalischen Füllung der Konjunktoren in monosyndetischen Koordinationen in den untersuchten Korpora. Gezeigt werden Konjunktoren mit einem Mindestvorkommen von 2.

dieser Funktion auf der Wortarten-Ebene ausgezeichnet sind. In Fällen der Verwendung von „bis“ wie in (27) enthält das STTS-Handbuch (Schiller u. a. 1999, 66) die klare Anweisungen, mit KON zu annotieren. Dies bedeutet im Fall der halbautomatischen Annotierung jedoch einen manuellen Korrekturschritt, dessen konsequente Durchführung gerne unterbleibt.

2.2.2.4 Monosyndetische Verwendung

Die lexikalische Füllung der Konjunktoren der monosyndetischen Koordinationen in NEGRA und TIGER ist in der Tabelle 2.9 auf Seite 25 abgebildet. Der Konjunktor „und“ ist mit über 85% erwartungsgemäss dominant. Er ist deshalb bei der Anzahl Konjunkte, welche durch monosyndetisches „und“ verknüpft werden, auch der bestimmende Faktor in der Gesamtverteilung in Tabelle 2.12 auf Seite 28, wie die folgende separate Auswertung für monosyndetisches „und“ über allen Korpora zeigt:

(28) Auflistung des Konjunktors „und“ (Mittelwert 3.4 mit 0.89):

3-teilig (2067, 75.0%), 4-teilig (468, 17.0%), 5-teilig (129, 4.7%), 6-teilig (47, 1.7%), 7-teilig (20, 0.7%), 8-teilig (13, 0.5%), 9-teilig (9, 0.3%), 10-teilig (2, 0.1%), 13-teilig (1, 0.0%), 11-teilig (1, 0.0%)

Der zweithäufigste Konjunktor „oder“ verknüpft etwas weniger Konjunkte pro Vorkommen:

(29) Auflistung des Konjunktors „oder“ (Mittelwert 3.28 mit Standardabweichung 0.62):
3-teilig (210, 79.2%), 4-teilig (38, 14.3%), 5-teilig (15, 5.7%), 7-teilig (1, 0.4%), 6-teilig (1, 0.4%)

Stellt sich die Frage, ob der dritthäufigste Konjunktor „sowie“ dieselben Eigenschaften hat. Für den Grammatik-Duden (Dudenredaktion 2005, §935) wird die Konjunktion „sowie“ als „Anreihung eines Nachtrags im Sinne von 'auch noch'“ sowie für die „Gliederung komplizierter Reihungen“ verwendet. Im GRAMMIS (IDS 2006) wird im Grammatischen Wörterbuch unter dem Eintrag zum Konjunktor „sowie“ folgende Funktionsbeschreibung gegeben: „'Sowie' dient vor allem der Gliederung in koordinativen und-Verknüpfungen mit drei und mehr Koordinaten [=Konjunkte] sowie der Vereindeutigung von koordinativen Verknüpfungen mit Koordinaten auf mehreren syntaktischen Hierarchieebenen.“ Die folgende Auflistung der Konjunktanzahl zeigt dann auch ein etwas anderes Bild, insbesondere die sonst häufigsten Fälle mit 3 Konjunkte sind prozentual weniger vertreten:

(30) Auflistung des Konjunktors „sowie“ (Mittelwert 3.64 mit Standardabweichung 1.29):
3-teilig (95, 65.1%), 4-teilig (33, 22.6%), 5-teilig (9, 6.2%), 6-teilig (4, 2.7%), 8-teilig (2, 1.4%), 11-teilig (2, 1.4%), 7-teilig (1, 0.7%)

Durch Auswertung auf der Ebene der unmittelbaren Tochterkonstituenten wird die Funktion von „sowie“ als koordinativer Verknüpfer von koordinativ Verknüpftem nicht richtig erfasst, wie das Beispiel (31) zeigt, wo die Binnenkoordinationen nicht direkt in den Tochterkonjunkten bestehen, sondern eher strukturell tiefer liegen.

(31) Darüber hinaus besteht das Forum auf [$_{CNP}$ [$_{NP}$ die Respektierung des Rechts auf Selbstbestimmung der Lateinamerikaner] , [$_{NP}$ die aktive Bekämpfung von [$_{CNP}$ Rassismus und Fremdenhaß] in den Industrienationen] sowie [$_{NP}$die Rückzahlung des Reichtums der Ersten Welt aus [$_{CNP}$ Sklaverei und ungerechtem Welthandel] an die Länder der Dritten Welt bis 12. Oktober]] . [N$_{9452}$]

Die Tabelle 2.15 auf der nächsten Seite, welche die Verteilung der 3 häufigsten Konjunktoren auf die Koordinationsklassen enthält, weist für „sowie“ im Vergleich zu „und“ bzw. „oder“ eine deutlich höhere monosyndetische Verwendung nach.

in %	Anzahl	in%	Koord.-Typ	Anzahl	in %
und	20993	100.0	syn	18125	86.3
			mono	2757	13.1
			x	111	0.5
oder	1797	100.0	syn	1505	83.8
			mono	265	14.7
			x	27	1.5
sowie	614	100.0	syn	450	73.3
			mono	146	23.8
			x	18	2.9

Tabelle 2.15: Verteilung der 3 häufigsten Konjunktoren über die Koordinationstypen in allen Korpora

2.2.2.5 Andere Koordinationstypen

In der Tabelle 2.9 auf Seite 25 zeigt sich noch ein kleiner Anteil von konjunktorhaltigen Koordinationen, welche weder syndetisch noch monosyndetisch sind. Die Verteilung der Funktionen der Tochterkonstituenten in der Tabellenzusammenstellung 2.16 auf der nächsten Seite für NEGRA und TIGER zeigt, dass viele verschiedene Varianten mit niedriger Häufigkeit vorkommen. Normalerweise liegt eine monosyndetische Grundstruktur vor, bei der zusätzlich nochmals Konjunkte weiter vorne zusammengefasst werden.

In (32a) ist eine typische Reihung zweier Teil-Koordinationen illustriert. Satz (32b) zeigt den zweithäufigsten Fall, bei dem die Paarformel „Kaffee und Kuchen“ durch „sowie“ integriert wird.

(32) a. [$_{CNP}$ [$_{NN-CJ}$ Skulpturen] [$_{KON-CD}$ und] [$_{NN-CJ}$ Keramiken], [$_{NN-CJ}$ Goldschmiedearbeiten] [$_{KON-CD}$ und] [$_{NN-CJ}$ Seidenmalereien]] sind auf dem Kunstmarkt zu sehen, [...] [N$_{12359}$]

b. Nach Angaben der Veranstalter sorgten [$_{CNP}$ [$_{NN-CJ}$ Spiele], [$_{NP-CJ}$ eine Tombola] [$_{KON-CD}$ sowie] [$_{NN-CJ}$ Kaffee] [$_{KON-CD}$ und] [$_{NN-CJ}$ Kuchen]] bei den Besuchern für Freude. [N$_{13307}$]

Die involvierten Konjunktorenfolgen sind für NEGRA in der Auflistung (33) und für TIGER in (34) aufgelistet. Die Vermutung, dass „sowie“ tendenziell als letzter Konjunktor erscheint, wird durch die Daten nicht richtig gestützt.

(33) Konjunktorfolgen in NEGRA:
„und und“ (24), „sowie und“ (7), „und sowie“ (7), „oder und“ (6), „und und und“ (4), „oder oder“ (2), „und oder“ (2), „ferner sowie“ (1), „sondern sowie“ (1), „sowie oder“ (1), „und doch“ (1)

(34) Konjunktorfolgen in TIGER:
„und und“ (35), „und sowie“ (7), „und oder“ (5), „und und und“ (4), „so-

NEGRA (total 58)

in %	Anzahl	Tochterfunktionen	kumulativ
32.8	19	CJ CD CJ CJ CD CJ	33
17.2	10	CJ CJ CD CJ CD CJ	50
8.6	5	CJ CJ CJ CD CJ CD CJ	59
5.2	3	CJ CD CJ CJ CJ CD CJ	64
3.4	2	CJ CD CJ CJ CD CJ CJ CD CJ	67
3.4	2	CJ CJ CD CJ CJ CD CJ	71
3.4	2	CJ CJ CJ CD CJ CJ CJ CD CJ	74
3.4	2	CJ CJ CJ CD CJ CJ CJ CJ CJ CD CJ	77

TIGER (total 71)

in %	Anzahl	Tochterfunktionen	kumulativ
33.8	24	CJ CD CJ CJ CD CJ	34
15.5	11	CJ CJ CD CJ CD CJ	49
8.5	6	CJ CJ CD CJ CJ CD CJ	58
5.6	4	CJ CD CD CJ	63
4.2	3	CJ CD CJ CJ CD CJ CJ	68
4.2	3	CJ CD CJ CJ CD CJ CJ CD CJ	72
2.8	2	CJ CD CJ CJ CJ CD CJ	75
2.8	2	CJ CJ CJ CD CJ CD CJ	77

Tabelle 2.16: Verteilung der Tochterfunktionen der Koordinationen vom Typ x in NEGRA und TIGER. Gezeigt werden die Fälle mit mindestens 2 Vorkommen. NEGRA hat 21 verschiedene Typen mit einem Type-Token-Verhältnis von 1:2.8. TIGER hat 24 Typen mit einem Type-Token-Verhältnis von 1:3.0.

wie und“ (3), „oder aber“ (2), „oder oder“ (2), „aber und“ (1), „oder und“ (1), „sowie sowie und“ (1), „und sowie und“ (1), „und und und und sowie und“ (1), „wie wie“ (1)

2.2.2.6 Spezialfälle

Die Abkürzung „usw.“ für das Mehrwortlexem „und so weiter“ tritt in NEGRA 7 Mal auf und wird konsistent als Adverb getaggt. Ebenfalls „etc.“, das in NEGRA nur 2 Mal und dafür in TIGER 15 Mal erscheint. Da es Konjunktor und Konjunktfunktion in einer Wortform vereinigt, stellt sich die Frage, wie es strukturell eingebaut werden soll. Wenn „usw.“ postnominal an ein einzelnes Nomen gehängt wird wie in (35a), ergibt sich keine Koordination. Sobald „usw.“ in einer Aufzählung erscheint, wird es als selbstständiges Konjunkt annotiert wie in (35b), das nicht einmal asyndetisch angebunden ist und auf Grund der Kategorienunverträglichkeit im Prinzip zu CO-Koordination führen sollte.

(35) a. [...] zwischen menschlichen Fähigkeiten in ihrem vollen Spektrum [$_{PP}$ [$_{APPR-AC}$ mit] [$_{NN-NK}$ Intuition] [$_{ADV-MNR}$[10] usw.]] und Fähigkeiten der Ökosysteme. [N$_{6919}$]

b. [$_{CNP}$ [$_{NP-CJ}$ Unvorhergesehene Reparaturen] , [$_{NP-CJ}$ nicht abgesprochene Preisgrenzen] , [$_{NP-CJ}$ Unerreichbarkeit des Auftraggebers] [$_{ADV-CJ}$ usw.]] führen zu Komplikationen, die bei richtigem Verhalten vermeidbar gewesen wären. [N$_{11568}$]

2.3 Morphemkoordination

Unter Morphemkoordinationen werden koordinierte Strukturen aufgefasst, bei denen mindestens ein Konjunkt aus einem Teilwort (Morphem) besteht, das in schriftlicher Form am Anfang oder am Schluss mit einem sogenannten Ergänzungsstrich markiert ist.

Die Funktion des Ergänzungsstrichs wird in der neuen Rechtschreibung (Drosdowski 2000) folgendermassen umschrieben:

> §98: Mit dem Ergänzungsstrich zeigt man an, daß in Zusammensetzungen oder Ableitungen einer Aufzählung ein gleicher Bestandteil ausgelassen wurde, der sinngemäß zu ergänzen ist.

Das sinngemässe Ergänzen, das sich immer auf sprachliches Material im Kotext stützt, ist ein wesentliches Merkmal und unterscheidet den Ergänzungsstrich von graphematischen Mitteln wie den Auslassungspunkten, welche etwas fragmentarisch Ausgedrücktes markieren, das typischerweise nicht durch Material aus dem Kotext zu vervollständigen ist.

Der Ergänzungsstrich als graphematisches Grenzsignal (Gallmann 1989) wird in der mündlichen Sprache intonatorisch nicht markiert. Typographisch unterscheidet sich der Ergänzungsstrich nicht vom Binde- und Trennungsstrich, welche alle

drei im Schriftsatz mit dem kürzesten Strich realisiert werden.[11] In Texten lassen sich Ergänzungstriche am Zeilenende nicht eindeutig von Trennungs- und Bindestrichen unterscheiden. Dies führt im NEGRA-Korpus dazu, dass in mehr als 60 Fällen Bindestrichkomposita und einige Trennungen in zwei Token aufgeteilt wurden. Die Schwierigkeiten, die im und aus dem Umgang mit nicht-trivialen Tokenisierungen entstehen, werden im Abschnitt 2.3.3 besprochen.

Im prototypischen Beispiel (1) auf Seite 9 haben wir es also mit einer syndetisch koordinierten Struktur (Konjunktor „und“) mit zwei Konjunkten zu tun. Das erste Konjunkt ist durch das Morphem „Polizei“ realisiert, das zwar als selbstständiges lexikalisches Morphem existiert, wegen des Ergänzungstrichs am Morphem-Ende aber im Text als unvollständig markiert ist. Ausgelassen ist „arbeiten“ aus dem nachfolgenden Konjunkt „Verwaltungsarbeiten“ – somit liegt eine Katalepse vor.

2.3.1 Morphemkoordination und Ellipsenverständnis

Wie etwa in Zifonun u. a. (1997, 2360) gehen viele Autoren grundsätzlich von einem symmetrischen Koordinationsschema aus, d.h. im Normalfall sind die beiden Konjunkte in der syntaktisch gleichen Klasse X:

$$[\text{Konjunkt}]_X \text{ Konjunktor } [\text{Konjunkt}]_X$$

In der Morphemkoordination werden nur in absoluten Ausnahmefällen keine vollständigen Wörter in einer Morphemkoordination vorkommen. Im ganzen NEGRA-Korpus finden sich drei Beispiele:[12]

(36) a. Außerdem sinke der Anteil der Lohn- und Gehaltsfortzahlung bei Krankheit an der *Bruttolohn- und -gehaltssumme* seit Jahren: [N_{3957}]

b. Die (Weiter-)Existenz und Minimalqualitäten des *Schienen-Nah- und -Regionalverkehrs* muß [...] [N_{957}]

c. [...] mit den künftigen *S-Bahn-Über- und -unterführungen* [...] [N_{16410}]

Somit sind Morphemkoordinationen fast ausschliesslich asymmetrische Koordinationen. Im Gegensatz dazu steht die Sichtweise, welche Ellipsen strukturell ergänzt. Dann verschwindet das Phänomen der Morphemkoordination auf der strukturellen Ebene, und es wird meist eine symmetrische Wortkoordination vorliegen. Wer die eigene Terminologie genau nimmt, sollte deshalb im Beispiel (1) auf Seite 9 nicht von kataleptischer Morphemkoordination sprechen, sondern von kataleptischer Wortkoordination. In den Beispielen (36) liegt nach elliptischer Sicht dann eine Wortkoordination vor, welche gleichzeitig analeptisch und kataleptisch ist.

[11] In typographisch wenig differenzierter Textrepräsentation kann noch das Minuszeichen vor Zahlen ein Verwechslungskandidat sein.

[12] Im TIGER-Korpus sogar nur eines.

2.3.2 Links- und rechtselliptische Formen

Auslassungen von Wortteilen sind nur entweder am Anfang oder am Ende von Wörtern möglich. Also nicht in der Mitte und nicht gleichzeitig am Anfang und am Ende:

(37) a. * Sie studierte sowohl Morphem-Koordinationsphänomene wie Wort-~~Koordinations~~beschränkungen.

b. * Sie studierte sowohl Morphem-~~Koordinations~~phänomene wie Wort-Koordinationsbeschränkungen.

c. * Sie studierte sowohl Satzkoordinationsphänomene wie -subordinations-intensiv.

d. Sie studierte sowohl Satzkoordinations- wie -subordinationsphänomene intensiv.

Warum funktionieren kataleptische Mittelellipsen wie in Beispiel (37a) oder analeptische Konstrukte wie in (37b) nicht? An den morphologischen Sollbruchstellen alleine kann es nicht liegen, denn an beiden Stellen sind Auslassungen eigentlich möglich. Werden hier strukturelle Beschränkungen verletzt? Oder sind diese Fälle eher aus Gründen der Sprachverarbeitung ausgeschlossen? Ein Rezipient kann in diesen Beispielen keine Lücke „hören“, welche noch semantisch ergänzt werden sollte. Anderseits ist ein Zuviel an Lücken wie in Beispiel (37c) ebenfalls nicht interpretierbar. Dies, obwohl gleichviel sprachliches Material vorhanden ist wie im akzeptablen Beispiel (37d).

Wie im Folgenden noch gezeigt wird, sind Linksellipsen immer analeptische Strukturen und Rechtsellipsen immer kataleptisch.

2.3.3 Morphemkoordination und TRUNC

Durch den Ergänzungsstrich, der als Teil der graphematischen Erscheinung eines elliptischen Wortes aufgefasst wird, lassen sich Rechtsellipsen in einem Korpus recht einfach einkreisen. In NEGRA gibt es 564 Token[13] in 510 Sätzen, welche mit einem Ergänzungsstrich enden und noch mindestens ein Zeichen davor enthalten. Die Beschreibung des STTS-Wortarten-Tags TRUNC lautet im STTS-Annotationshandbuch (Schiller u. a. 1999, 74):

> Mit TRUNC werden Wortteile bezeichnet, die mit einem Bindestrich enden, der einen Teil des nachfolgenden, mit *und*, *oder* verknüpften Wortes ersetzt.

Diese Definition beinhaltet 5 Komponenten:

[13] Die 61 Fälle, wo ein Trennungs- bzw. Bindestrich am Zeilenende fälschlicherweise zur Auftrennung eines Worts in 2 Token führte und eine Korrektur keinen Einfluss auf die syntaktische Struktur des Satzes hat, sind dabei schon stillschweigend korrigiert. Der umgekehrte Fall, dass ein Ergänzungsstrich als Trennungsstrich aufgefasst wird, ist mir nicht aufgefallen; die seltsame „Blut-, Schweiß-und-Tränen-Rede“ [N_{19057}] hat im Original die gleiche (falsche) Form.

- graphematische Auszeichnung: Ergänzungsstrich
- elliptische Kürzung: Mitverstandene Auslassung
- kataphorische Ergänzung: Mitverstandenes Material erscheint nach Kürzung
- morphologische Position: Erstglied (d.h. Rechtellipse)
- syntaktische Struktur: Koordination

Da im NEGRA die Wortarten eigentlich gemäss STTS kategorisiert sind, sollten alle kataleptischen Morpheme einfach zu finden sein, da sie und nur sie das Tag TRUNC tragen. Allerdings wurde jede dieser 5 Komponenten in der Annotierpraxis unterlaufen, wie im Folgenden diskutiert wird.

TRUNC bei Linksellipsen Im Original-NEGRA sind 12 Morpheme mit TRUNC ausgezeichnet, welche den Ergänzungsstrich am Anfang tragen. Das Markieren von Linksellipsen mit TRUNC wäre aus einer strukturellen Sicht durchaus eine Option. Ungünstig ist im NEGRA-Korpus, dass es keine einheitlich befolgte Annotationskonvention gibt, d.h. 10 Linksellipsen sind nicht mit TRUNC markiert.[14] Für die folgende Untersuchung wurden die 12 Fälle in ihre entsprechende Kategorie korrigiert.

Warum ist TRUNC überhaupt auf Erstglieder eingeschränkt? Beim Design des STTS-Tagsets hatte man für TRUNC die morphosyntaktische Generalisierbarkeit im Auge: Da im Deutschen viel Information in den Suffixen steckt, wurde die Klasse TRUNC von den andern lexikalischen Wortklassen zur Erhaltung von Endungsregularitäten abgesondert. Eine Bestimmung der morphologischen Merkmale wie Numerus oder Kasus ist bei einem gekürzten Wort ohne den Kontext nicht möglich. Dasselbe Problem tritt allerdings etwa auch bei Abkürzungen auf. Für diese Fälle wird im STTS-Annotationshandbuch (Schiller u. a. 1999, 9) jedoch festgehalten: „Es gibt kein eigenes Tag für Abkürzungen. Abgekürzte Wortformen werden generell so getaggt wie die ausgeschriebene Form.“ Im Gegensatz dazu werden im Münsteraner Tagset (Steiner 2003, 14) elliptische Wörter konsequent wie ihre vollständigen Formen getaggt, Abkürzungen hingegen erhalten auf der primären Ebene ein wortartenunspezifisches Abkürzungstag „Abk“, wobei dank der dort eingesetzten mehrstufigen Wortartenklassifikation auf der 2. Ebene noch eine der vollständigen Form entsprechende Klassifikation zugeschaltet wird.

TRUNC ohne Ergänzungsstrich In NEGRA werden 16 Token mit TRUNC ausgezeichnet, welche gar keinen Ergänzungsstrich beinhalten. Diese Fälle sind keineswegs krude Annotationsfehler, sie stehen alle im Zusammenhang mit proble-

[14] Dasselbe uneinheitliche Bild zeigt sich ebenfalls im TIGER-Korpus.

matischen Tokenisierungen bzw. Tokenisierungsfehlern, welche via Wortart kompensiert werden.[15]

(38) a. Der " Fraunhof " -Sonderorden [...]
ART $(TRUNC $(NN

[N_{15}]

b. Tante Emma-Laden , Bäcker und Metzger [...]
TRUNC NN $, NN KON NN

[N_{994}]

Hier wird TRUNC verwendet, um ein Morphem in das nachfolgende Wort zu inkorporieren, ohne dass in irgendeiner Form eine koordinative Verknüpfung vorliegt. Die Verwendung von paarigen doppelten Hochkommata wie in (38a) dient meist dazu, uneigentliche Rede (als Zitat-Ausdruck, Ironisierung usw.) im Text zu markieren. Wenn doppelte Hochkommata wie diese als satzintern und nicht als wortintern aufgefasst werden sollen, ergeben sich zwingend Schwierigkeiten.

Auch wenn das Leerzeichen im Quasi-Zitat „Tante Emma“ im Beispiel (38b) orthographisch mit Bindestrich hätte durchgekoppelt werden müssen, sind solche Konstruktionen bei Strassennamen wie „Albert Anker-Str.“ zulässig – und sie werden tendenziell häufiger bei der Verwendung von englischen Komposita.[16] Doppelte Hochkommata und Leerzeichen können selbstverständlich auch miteinander in einem Mehrwortlexem auftauchen, womit sich die Frage nach dem Status der einzelnen Teile sogar verschärft stellt.

Die *isolationistische* Tagging-Strategie für mehrteilige Wörter wie sie im STTS-Handbuch (Schiller u. a. 1999, 9) als Notlösung für Mehrwortlexeme empfohlen wird, muss im Beispiel (38a) den eigenen Annotierregeln folgend „Fraunhof“ als NE und „Tante“ als NN kategorisieren. Eine syntaktische Annotation, welche analog zur Phrasenkategorie „mehrteiliger Eigenname“ (multi-word proper noun), kurz MPN, auch ein mehrteiliges Substantiv erlauben würde, kurz MNN, könnte die funktionale Mehrdeutigkeit innerhalb von Nominalphrasen beseitigen und zu einer konsequenteren Umsetzung des isolationistischen Taggings führen.

[15] Tokenisierungen wie in (38a) sind nicht durchgängig so gemacht, es gibt mindestens 10 Wörter in NEGRA, wo die doppelten Anführungszeichen im Worttoken integriert sind; in 8 Fällen sind Klammern wie in den Beispielen (39) auf der nächsten Seite in das Wort integriert.

[16] Folgende Beispiele zeigen zwei unterschiedliche Auswege aus dem Tokenisierungs- bzw. Tagging-Notstand. In (ia) wird die Einheit „Wall Street-“ betont und als MPN zusammengefasst. In (ib), wo ein flache PP annotiert ist, erhält „Hop-“ wegen der durch den Bindestrich betonten Unvollständigkeit das Tag TRUNC, obwohl es hier gleich unvollständig ist wie „Hip“.

(i) a. [...] als Wall Street- Anwalt [...]
APPR NE NE NN

[N_{19910}]

b. [...] beim Sony Hip Hop- Festival [...]
APPRART NE NE TRUNC NN

[N_{3980}]

Eine *funktionsorientierte* Tagging-Strategie, welche ein Token nach seiner Rolle im Kontext kategorisiert, müsste hingegen für solche Fälle ein eigenes Wort-Tag einführen, das sich für solche abgesetzten Teilwörter verwenden lässt.

Im Münsteraner Tagset (Steiner 2004, 8) wird das Dilemma zwischen „isolierendem“ Tagging, das jedes Token möglichst nah an seiner Einzeldistribution kategorisiert und syntaktisch-funktionalem Tagging, welches über die Token-Grenze hinaus mehrteilige Wortgrössen zu identifizieren sucht, gelöst, indem sogenannte Metatags gesetzt werden können, welche Token zusammenfassen und in ihrer Funktion bestimmen. So würde ein Organisationsname wie „Bundesverband der Deutschen Industrie“ primär isolationistisch als „Bundesverband#Ng der#Db Deutschen#Aa Industrie#Ng“ getaggt. Darauf aufbauende Metatags markieren dann Beginn/Fortführung ($) und Ende (^) für den daraus zusammengesetzten Namen: „Bundesverband#Ng$Ne der#Db$Ne Deutschen#Aa$Ne Industrie#Ng^Ne“.

TRUNC bei Klammerausdrücken Die folgenden Beispiele haben im Gegensatz zu den vorangehenden eine Semantik, welche eine „und“-Beziehung ausdrücken.

(39) a. Die Kappellensanierer (innen) sind [...]
ART NN $(TRUNC $(VAFIN
[N_{3411}]

b. [...] des Alt- (*Ober*) Bürgermeisters [...]
ART TRUNC $(TRUNC $(NN
[N_{11694}]

Im Satz (39a) liegt eine Kurzform vor, welche eine der beiden in der offiziellen Rechtschreibung empfohlenen zulässigen Möglichkeiten darstellt, um die Doppelnennung „Die Kappellensanierer und Kappellensaniererinnen“ zu vermeiden. Die andere mögliche Kurzform müsste „Kappellensanierer/-innen“ mit Schräg- und Ergänzungsstrich lauten. Solche Kurzformen sind nur in der schriftliche Ausdrucksform möglich. An der Stelle, wo der Ergänzungsstrich auftaucht, kann mündlich keine Auslassung gemacht werden:[17]

(40) * Die Kappellensanierer und ~~Kappellensanierer~~-innen [...]

Nur wer TRUNC auch auf Letztglieder anwenden will und keinen Ergänzungsstrich verlangt, darf „innen“ damit markieren.

Beispiel (39b) belegt diesen Schreibstil, der kompakte Alternativformulierungen zulässt, welche auch vollständig im Innern des Wortes liegen können – etwas, das mit dem Schrägstrich unmöglich ist. In diesem Beispiel gibt es mit „Alt-“, „Ober“ und „Bürgermeister“ drei unvollständige Bestandteile: ein Erst-, Binnen- und Letztglied. Aufgrund der Annotationsbespiele im NEGRA-Korpus könnte hier ebensogut alles wie nichts mit TRUNC markiert werden.

[17] Der Duden erlaubt als Nebenform auch die Schreibweise ohne Ergänzungsstrich, d.h. „Kappellensanierer/innen“. In Anbetracht des Unterschieds im mündlichen Mitverstehen der Auslassung macht diese Variante noch Sinn.

Fazit: Um Tokenisierungsprobleme durch wortinterne Klammereinschübe kompensieren zu können bzw. unterschiedliche Tokenisierungsstrategien in einem Tagset zu unterstützen, wäre eine wortinterne Interpunktionskategorie nützlich. Weiter sollte analog zu abgetrennten Zifferngruppen wie in „079 787 88 88“ mit Hilfe von einem Funktionslabel analog zu NMC (*numerical component*) eine Teilwortkategorie geschaffen werden.

TRUNC ohne Rechtsellipse Nebst 64 Trennungsfehlern (61 davon sind wie in Fussnote 15 auf Seite 42 schon angesprochen stillschweigend korrigiert, da diese auf der syntaktischen Ebene keine Korrekturen mit sich ziehen) gibt es noch 17 Fälle, wo ein mit TRUNC markiertes Token keine Rechtsellipse darstellt, d.h. ein Binde- und kein Ergänzungsstrich vorliegt am Wortende. Klammern wie in Beispiel (41a) sind in 9 Fällen der Grund. In 2 Fällen sind Anführungszeichen wie in (41b) involviert, in 3 Fällen (englische) mehrteilige Komposita wie in Fussnoten-Beispiel ((ia) auf Seite 42).

(41) a. [...] die Ansiedlung von (Büro-) Arbeitsplätzen [...]
APPR $(TRUNC $(NN
[N_{3822}]

b. [...] , Mitglied des ANC- " Schattenkabinetts " , [...]
ART TRUNC $(NN $($(
[N_{5051}]

Fazit zur Verwendung von TRUNC Die Kategorie TRUNC verleitet im Kontext der Korpusannotation dazu, Tokenisierungsfehler zu kompensieren, welche insbesondere im Zusammenhang mit wortinternen Interpunktionszeichen entstehen und/oder an die spezifischen Mittel schriftlicher Sprache gebunden sind, wo sich im Gegensatz zur mündlichen Sprache sehr leicht Texte überlagern lassen mittels Klammerung und Schrägstrichen. Da diese Tokenisierungsfehler gerade häufig bei diesen schwierigen Fälle auftreten, sind entsprechende Korrekturtags wünschbar.

2.3.4 Rechtsellipsen und kataleptische Strukturen

Die Tabelle 2.17 auf der nächsten Seite gibt eine Übersicht, wie sich in NEGRA Wortformen mit einem Strichzeichen am Ende auf die Ebene der Wortarten, deren funktionaler Bestimmung und der Mutterkategorie verteilen. Die Verwendung der Rechtsellipsen zur direkten Koordination von nominalen Elemente zur koordinierten Nominalphrase (CNP) ist mit über 80% dominant. Direkt koordinierte Adjektivphrasen (CAP) sind mit knapp 6% vertreten. Weitere koordinierte Kategorien[18] wie CVP oder CAVP treten extrem selten elliptisch auf.

[18]Bei CO liegt eine fehlerhafte Annotation einer nominalen Morphemkoordination mit einem selbstständigen Ergänzungswort vor.

in %	Anzahl	Wortart	Funktion	Mutterkategorie	kumulativ
80.7	455	TRUNC	CJ	CNP	81
7.3	41	TRUNC	NK	NP	88
5.5	31	TRUNC	CJ	CAP	94
3.9	22	TRUNC	NK	PP	97
0.5	3	TRUNC	HD	VP	98
0.4	2	NE	PNC	MPN	98
0.4	2	TRUNC	PNC	MPN	99
0.2	1	NN	—	—	99
0.2	1	TRUNC	—	—	99
0.2	1	TRUNC	CJ	CAVP	99
0.2	1	TRUNC	CJ	CO	100
0.2	1	TRUNC	CJ	CVP	100
0.2	1	TRUNC	SB	S	100
0.2	1	XY	—	—	100
0.2	1	XY	NK	PP	100

Tabelle 2.17: Verteilung der 564 Token mit '-' am Wortende in NEGRA

Die homogenste Gruppe von Tabellenzeilen ist durch die Funktionsmarkierung CJ und die koordinierten Mutterkategorien gegeben, welche bei den hier betrachteten Token immer mit TRUNC korrelieren. Ausnahmen zu dieser Korrelation im Original-Korpus haben sich durchwegs als Fehlannotationen erwiesen.

Trotzdem enthält die Tabelle in über 7% der Fälle bei NP sowie in knapp 4% der Fälle bei PP gekürzte Wortformen, welche nicht unmittelbar Konjunkte einer koordinierten Phrase sind und somit keine prototypischen Fälle von elliptischer Wortkoordination darstellen. Im Folgenden werden diese Verhältnisse noch etwas genauer betrachtet.

Evaluation der Rechtsellipsen Wieviele rechtselliptische Token sind nun Bestandteil von Morphemkoordinationen? Um diese Frage und noch weitere interessierende Punkte exakt zu beantworten, wurden alle Vorkommen in vier Klassen kategorisiert und mit weiteren Buchstabenkodes für Untermerkmale wie folgt versehen:[19]

1. **Rechtselliptische Koordination**

   (a) Das Token ist Teil eines mehrteiligen Eigennamens.

   (b) Das Token ist Teil einer mehrteiligen Phrase, welche dann koordiniert ist.

[19]Die vollständige Liste, die halbautomatisch erstellt und manuell korrigiert wurde, ist elektronisch verfügbar. In diesem Abschnitt spreche ich nur einige Aspekte an, die sich daraus ergeben.

(c) Das Ergänzungstoken, d.h. dasjenige Token, welches das im kategorisierten Token fehlende Material enthält, ist Teil einer mehrteiligen Phrase, welche dann koordiniert ist.

(d) Das Ergänzungstoken hat ein explizites Determinativ-Element.

(e) Es liegt eigentlich keine Koordination vor.

(f) Das Ergänzungstoken besteht nur aus dem fehlenden Material.

(l) Das Ergänzungstoken ist selbst eine Linksellipse.

(x) Es liegt eine aus der Sicht der Koordination problematische syntaktische Annotation im NEGRA vor.

2 **Getrennt tokenisierte Komposita** ausgelöst durch

(a) „"“

(b) „)“

(c) (englisches) mehrteiliges Wort

3 **Textsortenspezifische Ideogramme** (In diesem Korpus treten nur im Zeitungsjournalismus übliche Autorkürzel auf.)

4 **Eindeutig falsche syntaktische Annotation**

Prototypische Fälle wie in Beispiel (1) auf Seite 9 erhalten somit Kode 1, das Beispiel (36) auf Seite 39) den Kode 1l.

Der Fall 1a ist im komplexen Beispiel (42a) illustriert, wo mehrteilige Eigennamen zur Konstituentenstruktur (MPN, *multi-lexeme proper noun*) zusammengefasst werden. Allerdings zeigt dieses Beispiel gleich das damit verbundene Problem: Während MPN bei der Rechtsellipse „Led Zeppelin-“ effektiv einen Eigennamen als Mehrwortlexem aus den Einzeltoken konstituiert, wird bei „Jimi Hendrix-Fans“ ein im Kern normales Substantiv zu MPN gemacht, obwohl eigentlich nur „Jimi Hendrix-“ als Teilbestandteil einen Eigennamen darstellt und der Kern „Fans“ ein normales Substantiv ist. Bei „Extreme-Jüngern“ im gleichen Satz liegt eigentlich dieselbe morphologische Konfiguration vor, aber in diesem Fall zählt der Ausdruck als normales Substantiv. Die in TIGER gemachte Tagset-Erweiterung NNE für Substantiv-Komposita, deren Erstglied aus einem Eigennamen und Letztglied aus einem normalen Substantiv besteht, würde in diesem Fall nur bedingt weiterhelfen, da es sich insgesamt um einen Mehrwortausdruck und kein Einzelwort handelt.

(42) a. [...] , weil das Powerpack nun ein breites Spektrum von Musikfans locker anspricht, von den altgedienten [$_{CNP}$ [$_{MPN-CJ}$ [$_{NE-PNC}$ Led] [$_{TRUNC-PNC}$ Zeppelin-]] , [$_{MPN-CJ}$ [$_{NE-PNC}$ Black] [$_{TRUNC-PNC}$ Sabbath-]] [$_{KON-CD}$ und] [$_{MPN-CJ}$ [$_{NE-PNC}$ Jimi] [$_{NN-PNC}$ Hendrix-Fans]]] bis zu den jungen [$_{CNP}$ [$_{TRUNC-CJ}$ Nirvana-] und [$_{NN-CJ}$ Extreme-Jüngern]] oder den Fusion-Fetischisten um Living Colour und Konsorten. [N_{252}]

b. [$_{CVP}$ [$_{VP-CJ}$ [$_{PTKNEG-NG}$ Nicht] [$_{TRUNC-HD}$ aus-]] , sondern [$_{VVINF-CJ}$ aufsteigen]] wollen Bernhard Mertens und Peter Weißenseel von der CDU. [N_{17278}]

c. Ob im Falle [$_{CNP}$[$_{NP-CJ}$ der Meister-] und [$_{NP-CJ}$ der Dürerschule] oder [$_{NP-CJ}$ des Höchster Kinderhauses]] - überall spiegelt sich der geringe Einfluß auf die Römerspitze wider. [N_{6611}]

Beim Fall 1b im Beispiel (42b) entsteht durch die kontrastierende Negationspartikel „nicht“ ein komplexes Erstglied, während das Ergänzungselement einfach bleibt, da „sondern“ als Konjunktor behandelt wird[20]. Der Fall 1bc tritt wie im Beispiel (42c) bei Koordinationen auf, wo der Artikel auch in folgenden Konjunkten wiederholt wird.

Die Annotationskonventionen in NEGRA verlangen für Koordinationen eine eigene Phrasenkategorie[21], in der alle koordinierten Elemente inklusive lexikalischer Koordinationsmittel auf einer Ebene eingeschlossen sind. Dies ist auch für Morphemkoordinationen so:

(43) [...] [$_{PP}$ [$_{APPR}$ von] [$_{ART}$ den] [$_{ADJA}$ übrigen] [$_{CNP-NK}$ [$_{TRUNC-CJ}$ Polizei-] [$_{KON-CD}$ und] [$_{NN-CJ}$ Verwaltungsarbeiten]]] [N_{168}]

Das Beispiel zeigt auch, dass in Präpositionalphrasen das direkt von der Präposition abhängige Material auf der gleichen Ebene flach annotiert wird. Dies erklärt, warum nominale Elemente in der Tabelle 2.17 auf Seite 45 überhaupt die Mutterkategorie PP haben können.

Aus der Tabelle 2.18 auf der nächsten Seite können alle potentiell in NEGRA vorhandenen echten Morphemkoordinationen abgelesen werden, welche nicht über die Koordination von Mehrwortlexemen hinausgehen: Sie haben entweder den Typ 1 und eine koordinierte Mutterkategorie oder Typ 1a und eine koordinierte Grossmutterkategorie. Die Auflistung der Verteilung der echten und unproblematischen Morphemkoordinationen ergibt folgendes Bild:

(44) CNP (426, 92.6%), CAP (30, 6.5%), MPN (2, 0.4%), CVP (1, 0.2%), CAVP (1, 0.2%)

Eine Übersicht über alle Fälle der Kategorie 2 bis 4 gibt Tabelle 2.19 auf der nächsten Seite. Darin sind auch alle nicht mit TRUNC getaggten Token enthalten, welche eine Rechtsellipse aufweisen.

CNP und Morphemkoordination Nominale Morphemkoordinationen sind absolut dominant. Die Tabelle 2.20 auf Seite 49 zeigt die Verteilung des strukturellen Aufbaus innerhalb der CNP. Mit 90% sind syndetische Koordinationen mit zwei

[20] Somit wird „nicht... sondern“ in Übereinstimmung mit Pasch (2003, 471) nicht als paariger Konjunktor angesetzt, wie es traditionell manchmal gemacht wurde.

[21] Im Gegensatz zu vielen grammatischen Theorien ist Phrasalität in NEGRA nicht auf maximale Konstituenten beschränkt.

in %	Anzahl	Mutterkategorie	Typ	kumulativ
79.3	426	CNP	1	79
5.6	30	CAP	1	85
4.1	22	NP	1bc	89
3.0	16	CNP	1c	92
1.9	10	PP	1bce	94
1.5	8	CNP	1x	95
1.3	7	PP	1bc	97
1.3	7	NP	1b	98
0.4	2	VP	1b	98
0.4	2	MPN	1a	99
0.4	2	CNP	1cx	99
0.2	1	VP	1bx	99
0.2	1	S	1ex	100
0.2	1	NP	1e	100
0.2	1	CVP	1	100
0.2	1	CAVP	1	100

Tabelle 2.18: Verteilung der 537 Token vom Typ 1 in der Evaluation der Rechtsellipsen in NEGRA

in %	Anzahl	Wortart	Typ	kumulativ
33.3	9	TRUNC	2b	33
18.5	5	TRUNC	4	52
11.1	3	TRUNC	2	63
7.4	2	NE	2c	70
7.4	2	TRUNC	2a	78
7.4	2	TRUNC	4a	85
3.7	1	NN	3	89
3.7	1	TRUNC	2c	92
3.7	1	XY	3	96
3.7	1	XY	4	100

Tabelle 2.19: Verteilung der 27 Token vom Typ 2 bis 4 in der Evaluation der Rechtsellipsen in NEGRA

in %	Anzahl	Konjunktor	Konjunkte
90.1	384	1	2
5.4	23	1	3
2.8	12	1	4
0.9	4	1	6
0.5	2	2	3
0.2	1	2	2

Tabelle 2.20: Verteilung der Konjunkt- und Konjunktoranzahl in den echten CNP-Morphemkoordinationen der 426 TRUNC-Token in NEGRA

in %	Anzahl	Tochterkonstituenten
88.6	359	TRUNC KON NN
3.2	13	TRUNC APPR NN
2.5	10	TRUNC KON NE
2.2	9	TRUNC TRUNC KON NN
1.0	4	TRUNC TRUNC TRUNC KON NN
0.7	3	NN TRUNC KON NN
0.2	1	TRUNC TRUNC TRUNC TRUNC NN KON NN
0.2	1	TRUNC NN KON NN
0.2	1	TRUNC KON TRUNC
0.2	1	TRUNC KON PIS
0.2	1	NP TRUNC KON NN
0.2	1	KON TRUNC TRUNC KON NN
0.2	1	KON TRUNC KON NN

Tabelle 2.21: Verteilung der Tochterkonstituenten der 405 CNP-Morphemkoordinationen in NEGRA

Konjunkten die typische Konstruktion. Die monosyndetischen decken nochmals etwa 9% ab. Die 2 Vorkommen mit 2 Konjunktoren und 3 Konjunkten stammen aus Beispielsatz (45), wo der zweiteilige Konjunktor „weder-noch“ monosyndetisch und flach mit 3 Konjunkten verknüpft wird. Das andere Einzelvorkommen ist eine normale Verwendung dieses Konjunktors.

(45) [...] fand die Polizei [$_{CNP}$ weder Beweis-, Bekleidungs- noch Schmuckstücke]. [N_{16921}]

Für eine strukturelle Betrachtung ist der lineare Bau einer Morphemkoordination wichtig. Die Tabelle 2.21 stellt die Tochterkonstituenten der Phrasen zusammen, wobei die Kommas nicht aufgeführt sind. Rechtselliptische Token treten linksperipher auf – abgesehen von wenigen Ausnahmen, welche mässig akzeptable und nicht ganz leicht zu interpretierende Konstruktionen darstellen: In (46a) besteht die Tendenz „einem Stück“ nur auf „Apfelstrudel“ zu beziehen; in (46b) scheint

Satz	Tochterkonstituenten
7333	TRUNC KON NN KON TRUNC KON NN
9187	NN TRUNC TRUNC KON NN TRUNC KON NN NP KON NN
9487	TRUNC KON NN NP
17351	TRUNC KON NN NP KON NP
19096	TRUNC TRUNC KON TRUNC KON NN

Tabelle 2.22: Die Tochterkonstituenten der problematischen CNP-Morphemkoordinationen vom Typ 1x in NEGRA

mir der Skopus von „die“ entgegen der Annotation auf „Staaten“ beschränkt.[22]

(46) a. [...] konnten sich Mütter und Väter mit einer Tasse Kaffee und einem Stück [$_{CNP}$ Apfelstrudel, Marmor- oder Käsekuchen] stärken. [N_{11031}]

b. [...] die Wiederkehr [...] des Nationalismus, „der die [$_{CNP}$ Staaten, Völker- und Sozialgemeinschaften] explodieren läßt“ [...] [N_{2412}]

Man könnte hier einen Effekt konstatieren, der verhindert, dass Determinative über ein nicht-elliptisches Wort hinaus den Skopus in eine nachgeordnete koordinierte Rechtsellipse gewinnen. So wäre etwa die Phrase „mit einem Stück Apfelstrudel, Marmorcake oder Käsekuchen stärken“ in (46a) eindeutiger.

Das Ergänzungswort stellt dabei immer das erste nachfolgend auftretende, nicht-rechtselliptische nominale Wort dar. Nur in zwei Fällen steht zwischen einer Rechtsellipse und dem Ergänzungswort kein lexikalischer Konjunktor: In (47a) ist eine sowohl formale wie inhaltlich stark motivierte Reihung involviert; Beispiel (47b) ist schwierig zu interpretieren (die seltsame Setzung der Anführungszeichen weist ebenfalls darauf hin).

(47) a. Während bei den Jungs [$_{CNP}$ A-, B-, C-, D-, E-Jugend und „Minis“] ihre Klassensieger ermitteln [...] [N_{17698}]

b. [...] wie das Verhältnis der Schüler an der IGS aussehen soll, die von der „Grundschule für [$_{CNP}$Haupt-, Realschule und Gymnasium] empfohlen“ wurden. [N_{10632}]

In den Phrasen des Typs „TRUNC APPR NN“ übernimmt das Wort „bis“, obwohl es als Präposition getaggt ist, die Rolle des Konjunktors.

In der Tabelle 2.18 auf Seite 48 gibt es noch 8 Token in 5 CNP, welche als problematisch (Typ 1x) aufgeführt sind. Die in NEGRA annotierte komplexe Struktur kann der Tabelle 2.22 entnommen werden. Es sind Konstruktionen, welche keine monosyndetische Struktur aufweisen oder wo in einer mehrteiligen syndetischen Struktur der Skopus der Konjunktoren unterschiedlich ist.

(48) a. Zwei Experten werden die rund 1000 [$_{CNP}$ Tenor- und Baßsänger sowie Sopran- und Altsängerinnen] begutachten, [...] [N_{7333}]

[22] Eine andere Lesart kann entstehen, wenn man „Staaten“ entgegen der Verschriftlichung als „Staaten-~~gemeinschaft~~“ auffasst.

b. Gezeigt werden [$_{CNP}$ Mountainbikes, Alltags-, Reise- und Liegeräder, Lasten- und Transporträder, ein Behinderten-Rad und Second-Hand-Räder]. [N$_{9187}$]

c. [...]: [$_{CNP}$ Schmuck- und Duftschachteln, Gießkeramiken und [$_{NP}$ Broschen mit hellenistischen Motiven]] . [N$_{9487}$]

d. [$_{CNP}$Energie- und Wassersparen, regenerative Energien und die Förderprogramme des Bundes [...]] sind dieses Mal das Thema. [N$_{17351}$]

e. Die [...] Umsatzrendite des Unternehmens mit den Sparten [$_{CNP}$ Medizin-, Sicherheits- sowie Luft- und Raumfahrttechnik] ist [...] [N$_{19096}$]

Ausser im Satz (48e) ergeben sich wie von selbst linksperiphere syndetische Strukturen, wenn man die Morphemkoordinationen je zu eigenen CNP zusammenfasst. Konstruktionen der Form [$_{CNP}$ ... [$_{CNP}$ ...] ...] sind keineswegs „verbotene“ Annotationen – es gibt im Gegenteil 43 Sätze, welche solche verschachtelten CNP enthalten. Aus Gründen der Generalisierbarkeit der Konstruktionen ist es deshalb wünschenswert, Annotationen wie in Tabelle 2.22 auf der vorherigen Seite auszuschliessen – getriggert durch ihre Struktur, welche weder asyndetisch, monosyndetisch noch syndetisch ist.

Wenn im Satz (48e) die Morphemkoordination „Luft- und Raumfahrttechnik“ gebildet wird, liegt eigentlich ein Fall der Kategorie 1c vor. Die Situation ist jedoch insofern speziell, weil im gleichen Wort mit „-technik“ und „-fahrttechnik“ zwei verschiedene Ellipsen-Ergänzungen vorliegen.

CAP und Morphemkoordination Von den 29 TRUNC in CAP-Konstruktionen sind fast alle zweigliedrig und syndetisch wie in Beispiel (49a), nur eines ist dreigliedrig und monosyndetisch. Die Übersicht in Tabelle 2.23 auf der nächsten Seite gibt die Verteilung der Phrasenkomponenten an. Auch hier wird die mit APPR markierte Präposition „bis“ als Konjunktor betrachtet. Im Erstglied sind dabei durchgehend Kardinalzahlen zu finden wie im Beispiel (49b).

Im Vergleich zu den CNP sind die CAP-Morphemkoordinationen einfach und homogen. Koordinierte ADJD wie im Beispiel (49c) sind mit 22% etwas untervertreten, wenn man von insgesamt 7081 ADJD und 20590 ADJA im NEGRA-Korpus ausgeht.

(49) a. [...]werden paritätische [$_{CAP}$ ressort- und aufgabenbezogene] Strukturgruppen eingesetzt [...] [N$_{922}$]

b. in den gemischten, [$_{CAP}$ vier- bis fünfköpfigen] Teams [...] [N$_{4026}$]

c. [...] konkrete Ziele [$_{CAP}$ mittel- und langfristig] zu formulieren, [...] [N$_{17985}$]

Rechtsellipsen in syntaktisch komplexer Umgebung Wie man aus der Tabelle 2.18 auf Seite 48 ablesen kann, gibt es über 60 unproblematische Fälle (d.h.

in %	Anzahl	Tochterkonstituenten
48.3	14	TRUNC KON ADJA
24.1	7	TRUNC KON ADJD
24.1	7	TRUNC APPR ADJA
3.4	1	TRUNC TRUNC KON ADJA

Tabelle 2.23: Verteilung der 29 rechtselliptischen CAP- Morphemkoordinationen in NEGRA

solche ohne x), wo eine Rechtsellipse nicht auf der Wortebene dem Ergänzungstoken nebengeordnet ist. Sie kann zu einer komplexen Struktur (Typ 1b) erweitert sein – in Satz (50a) etwa mit einem eigenen Begleiter. Oder das Ergänzungswort ist erweitert koordiniert (Typ 1c) wie in Satz (50b), wo ein attributives Adjektiv das Ergänzungswort noch modifiziert[23]. Es können zudem beide Glieder gleichzeitig erweitert auftreten (Typ 1bc) wie in Satz (50c), wo das Ergänzungstoken wegen der Partitivkonstruktion nicht einmal im syntaktischen Kopf erscheint. In Satz (50d) stecken sowohl Ellipse wie Ergänzung im postnominalen Genitiv zweier koordinierter PP.

(50) a. [...] für [$_{CNP}$[$_{NP}$ ein Touren-] , [$_{NP}$ ein Renn-] oder [$_{NN}$ Trekking-Rad]], [...] [N$_{2841}$]

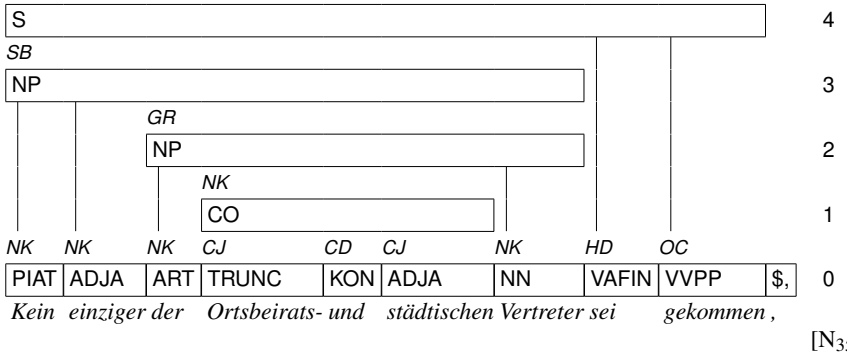
b. [...] 400 000 Tonnen [$_{CNP}$ [$_{TRUNC}$ Edel- und [$_{NP}$ nichtrostender Spezialstahl]] [...] [N$_{5862}$]

c. Allein [$_{CNP}$[$_{NP}$ die Ober-] und [$_{NP}$ Teile [$_{NP}$ der Mittelschicht]]] profitierten [...] [N$_{6185}$]

d. Die Mittel dafür sollen aber [$_{CPP}$ [$_{PP}$ nicht etwa aus dem Etat [$_{NP-GR}$ des Umwelt-] , sondern [$_{PP}$ aus dem [$_{NP-GR}$ des Entwicklungsministers]]] fließen. [N$_{20053}$]

[23] Eine alternative Annotation zu solchen Konstruktionen, wo das trunkierte Nomen mit dem nachfolgenden Adjektiv als CO-Koordination betrachtet wird, findet sich in folgendem Satz:

(i)



[N$_{3564}$]

Damit sollte wohl verhindert werden, dass die Morphemkoordination über eine NP-Grenze gehen soll. Dabei geht vergessen, dass „Ortsbeirats-“ sich semantisch mit „Vertreter“ verknüpfen sollte. Für mich klar eine Fehlannotation.

in %	Anzahl	Mutter	Funktion	Grossmutter	Tochterkonstituenten
31.6	12	NP	CJ	CNP	ART TRUNC
7.9	3	NP	CJ	CNP	ART ADJA TRUNC
7.9	3	NP	CJ	CNP	PIAT TRUNC
5.3	2	NP	CJ	CNP	ADJA TRUNC
5.3	2	NP	CJ	CNP	AP TRUNC
5.3	2	PP	CJ	CPP	APPR ART TRUNC
5.3	2	PP	MNR	NP	APPR TRUNC
5.3	2	VP	CJ	CVP	PTKNEG TRUNC
2.6	1	NP	CJ	CNP	ART AP TRUNC
2.6	1	NP	CJ	CNP	ART CARD TRUNC
2.6	1	NP	CJ	CNP	AVP TRUNC
2.6	1	NP	CJ	CNP	CARD TRUNC
2.6	1	NP	GR	NP	ART ADJA TRUNC
2.6	1	NP	GR	PP	ART TRUNC
2.6	1	NP	OA	S	ART TRUNC
2.6	1	PP	CJ	CPP	APPRART TRUNC
2.6	1	PP	MNR	NP	APPRART TRUNC
2.6	1	PP	MO	VP	APPR CARD TRUNC

Tabelle 2.24: Verteilung der 38 Fälle von Morphemkoordination vom Typ 1b in NEGRA

Bei über der Hälfte aller 64 Fälle liegt der Typ 1bc vor, wie folgende Auflistung zeigt: 1bc (60.9%), 1c (25.0%), 1b (14.1%) .

Die Tabelle 2.24 zeigt in der Übersicht die Struktur und Funktionalisierung der indirekt koordinierten Rechtsellipsen des Typs 1b(c). Determinative und/oder attributive Erweiterungen innerhalb von NP sind dabei am häufigsten und erkennbar an der Spalten-Kombination NP/CJ/CNP.

Rechtsellipsen in nicht-koordinativen Verknüpfungen Neben diesen koordinativ verknüpften Rechtsellipsen gibt es noch solche vom Typ e, welche ohne Koordination auftreten. Wie aus der Tabelle 2.18 auf Seite 48 entnommen werden kann, betrifft dies fast nur PP, wobei die Präposition „von“ eine besonders starke induzierende Wirkung hat und in 8 von 10 Fällen bei den Rechtsellipsen auftaucht. Sätze (51a) und (51c) belegen diese Fälle, wo typischerweise auch das Ergänzungstoken in einer PP vorkommt. Wie (51b) zeigt, ist dies aber nicht zwingend – und auch nicht auf Satzglieder gleicher Funktion beschränkt, da hier ein Subjekt die Ellipse eines Modifikators ergänzt. Wie (51d) zeigt, ist jedoch die Kontaktstellung relevant, d.h. die unmittelbare lineare Abfolge der betroffenen Satzglieder.

(51) a. [...] [$_{NP}$ Umwandlung [$_{PP-PG}$ von Miet-] [$_{PP-MNR}$ in Eigentumswohnungen]] [...] [N$_{4752}$]

in %	Anzahl	Mutterkategorie	Tochterkonstituenten
36.8	14	NP	ART TRUNC
10.5	4	NP	ART ADJA TRUNC
7.9	3	NP	PIAT TRUNC
5.3	2	NP	ADJA TRUNC
5.3	2	NP	AP TRUNC
5.3	2	PP	APPRART TRUNC
5.3	2	PP	APPR ART TRUNC
5.3	2	PP	APPR TRUNC
5.3	2	VP	PTKNEG TRUNC
2.6	1	NP	ART AP TRUNC
2.6	1	NP	ART CARD TRUNC
2.6	1	NP	AVP TRUNC
2.6	1	NP	CARD TRUNC
2.6	1	PP	APPR CARD TRUNC

Tabelle 2.25: Verteilung der 38 rechtselliptischen Morphemkoordinationen vom Typ 1 mit Untertyp b in NEGRA

b. [...] erfolgt [$_{PP-MO}$ nach der Vorrunden-] [$_{NP-SB}$ eine interne Mannschaftsumstellung] . [N_{5354}]

c. [$_{PP-MO}$Von der E-] [$_{PP-MO}$ bis zur A-Jugend] ist die FG 02 in allen Altersklassen mit Mannschaften vertreten. [...] [N_{11120}]

d. *? Eine interne Mannschafts- erfolgt nach der Vorrundenumstellung [...]

Der Satz (51c) mit der annotierten doppelten Vorfeldbesetzung[24] steht stellvertretend für die Schwierigkeit beim Behandeln von „bis“ in Kombination mit einer nachfolgenden Präposition. Wie Beispiel (52) zeigt, ist die Annotation in diesem Punkt nicht einheitlich.

(52) [$_{CPP}$[$_{PP}$ Vom einstigen VEB-Generaldirektor] [$_{PP}$ bis zum Pförtner]] habe man ihnen geglaubt [...] [N_{6151}]

Die beschriebenen Sprach-Daten zeigen, dass die Auslassung von Wortteilen keineswegs strikt auf die Koordination auf der Wortebene beschränkt ist, auch wenn sie damit prototypisch auftritt. Sie kann in seltenen Fällen auch in nicht-koordinativen Kontexten erscheinen, wo mehrere Konstituenten des gleichen Typs adjazent folgen.

2.3.5 Linksellipsen und analeptische Strukturen

Analeptische Strukturen sind viel seltener im Deutschen. Von 47 Token, welche im NEGRA-Korpus mit dem Zeichen „-“ beginnen, sind zudem nur ganze 24 linksel-

[24] Müller (2003) enthält eine umfangreiche Datensammlung zu diesem Thema.

in %	Anzahl	Tochterkonstituenten	kumulativ
66.7	16	NN KON NN^{l}	67
12.5	3	TRUNC KON NN^{l}	79
8.3	2	NN NN KON NN^{l}	88
4.2	1	NN NN^{l} KON NN	92
4.2	1	NE KON NE^{l}	96
4.2	1	ADJA KON ADJA^{l}	100

Tabelle 2.26: Verteilung der 47 linkselliptischen Morphemkoordinationen der Kategorie 1 in NEGRA. Die linkselliptischen Token sind mit hochgestelltem ^{l} markiert.

liptische Morphemkoordinationen. Wie bei den Rechtsellipsen wurde eine manuell validierte und kategorisierte Liste erzeugt, wobei die gleichen Kategorien, aber leicht angepasste Merkmale verwendet wurden:

1. **Linkselliptische Koordination**
2. **Getrennt tokenisierte Komposita** ausgelöst durch
   (a) „"“
   (b) „)“
   (c) (englisches) mehrteiliges Wort
   (d) „/“
3. **Textsortenspezifische Ideogramme** (Autorenkürzel)

In der Tabelle 2.26 sind die Tochterkonstituenten der Linksellipsen (ohne Interpunktion) zusammengestellt, wobei die einzelnen Ellipsen jeweils mit hochgestelltem ^{l} markiert sind. Das Ergänzungswort findet sich durchwegs links von der Ellipse im nächststehenden kategoriengleichen Wort.

Der einzige Fall aus Beispiel (53), wo mehr als eine Linksellipse in einer Phrase vorkommt (NN NN^{l} KON NN^{l}), ist dabei durch die beiden Einträge „NN NN KON NN^{l}“ bzw. „NN NN^{l} KON NN“ repräsentiert.

(53) [...] wie man das in südamerikanischen Maskenbräuchen, -festen und -tänzen seit der Eroberung durch die Europäer eingeführt hat [...] [N$_{206}$]

Beispiel (54a) illustriert den häufigsten Typ 1 der linkselliptischen Morphemkoordinationen mit 2 Konjunkten. Bei mehr als 2 Konjunkten wie in Beispiel (54b) erscheint mir diese Konstruktion schon schwieriger für den Rezipienten. Auch in (54c) ist die Verschriftlichung mi dem Ergänzungsstrich (und dem vorangehenden Bindestrich) für die korrekte Interpretation notwendig.

(54) a. dazu muß ein Bundes-ÖPNV-Gesetz die bundesrechtlichen [$_{CNP}$ [$_{NN-CJ}$ Rahmenbedingungen] und [$_{NN-CJ}$ -verpflichtungen]] [...] [N$_{956}$]

in %	Anzahl	Wortart	Typ	kumulativ
39.1	9	NN	2a	39
17.4	4	NN	2ac	56
13.0	3	NN	2b	70
4.3	1	NE	3	74
4.3	1	NN	2	78
4.3	1	NN	3	82
4.3	1	TRUNC	2a	87
4.3	1	TRUNC	2ac	91
4.3	1	TRUNC	2d	95
4.3	1	XY	3	100

Tabelle 2.27: Verteilung der 23 linkselliptischen Token vom Typ 2 bis 3 in NEGRA

b. In den letzten sechs Jahren sei hier deshalb über eine Million Mark für die Sanierung von [$_{CNP}$ [$_{NN-CJ}$ Hochbehältern] , [$_{NN-CJ}$ Tiefbrunnen] und [$_{NN-CJ}$ -pumpen]] sowie neue Rohre ausgegeben worden. [N_{18266}]

c. Hilfe zur Selbsthilfe, heißt denn auch das Motto, und das Angebot richtet sich an alle [$_{CAP}$ [$_{ADJA}$ bi-nationalen] und [$_{ADJA}$ -kulturellen] Paare. [N_{8799}]

Wie in der Tabelle 2.27 ersichtlich, sind es hauptsächlich durch doppelte Anführungszeichen entstandene, problematische Tokenisierungen, welche mit Linksellipsen konkurrenzieren.

2.4 Wortkoordination

2.4.1 Generelles

Während sich Morphemkoordination am graphematischen Kriterium des Ergänzungsstrichs festmachen lässt, sind bei der Wortkoordination folgende Punkte zu bedenken:

- Einzelne Wörter können die gleichen syntaktischen Funktionen wie Wortgruppen oder Phrasen tragen, und deshalb gibt es viele Fälle, wo einzelne Wörter mit Wortgruppen oder Phrasen gemischt koordiniert sind.

- Da sich Phrasen über ihre syntaktische Funktion konstituieren, welche wie vorhin angesprochen sowohl durch Einzelwörter wie Wortgruppen erfüllt werden kann, lassen sich viele Wortkoordinationen in funktionaler Sicht als Phrasenkoordinationen auffassen.

- Bei gewissen Partikeln wie z.B. „auch“ ist die Annotationskonvention in NEGRA, dass durchgängig mit ADV annotiert werden muss. Auch dann, wenn es Teil des mehrteiligen Konjunktors „sowohl – als auch“ oder „oder auch“

ist. Da das Annotationsmodell keine komplexe Konjunktorkategorie kennt, wie man sie sich analog zu den komplexen Personennamen MPN vorstellen kann, muss „auch“ als Modifikator zum nachfolgenden Konjunkt gezogen werden wie im Beispiel (55a). Dieses wird dadurch auf jeden Fall zu einer Wortgruppe.

(55) a. [$_{CNP}$[$_{KON}$ Sowohl] [$_{NN}$ Gesangssolisten] [$_{KON}$ als] [$_{NP}$ [$_{ADV}$ auch] [$_{NN}$ Pianisten]]] sind Schülerinnen und Schüler des Gagern-Gymnasiums. [N_{2941}]

Die einfachste Form von Wortkoordinationen bilden die nicht-phrasalen Formen wie koordinierte Präpositionen (CAC), koordinierte (subordinierende) Konjunktionen (CCP) sowie koordinierte Infinitive mit „zu“ (CVZ). Da diese Formen aber sehr selten sind, werde ich sie erst am Schluss besprechen.

Für die folgenden empirischen Untersuchungen gehe ich davon aus, dass Wortkoordination vorliegt, wenn alle Konjunkte aus exakt einem Wort oder Teilwort bestehen.

Die Tabelle 2.43 auf Seite 101 zeigt die Verteilung aller koordinierten Phrasen, welche nur Konjunkte aus Wörtern enthalten (W), unter allen Konjunkten mindestens eines enthalten, das nur aus einem Wort besteht (WP), bzw. denjenigen, wo kein Konjunkt aus einem Einzelwort besteht (P). Bei allen Phrasen ausser den CNP gibt es eine eindeutige Tendenz zu Wort- oder Phrasenkoordination. Wenn man sich nicht für die Koordinationskategorien interessiert, ergibt sich folgende Verteilung für NEGRA: P (5466, 55.0%), W (3444, 34.6%), WP (1031, 10.4%).

2.4.2 CNP

Wie in Tabelle 2.43 auf Seite 101 ersichtlich, bilden CNP-Wortkoordinationen in NEGRA mit 2577 Vorkommen den grössten Anteil aller Wortkoordinationen. Wenn wir nur die lexikalischen Konjunkte betrachten und die Konjunktoren ausser Acht lassen, ergeben sich insgesamt 2291 verschiedene Types mit einem fast ausgeglichenen Type-Token-Verhältnis von 1:1.1 . Die exakte Verteilung der Vorkommenshäufigkeit ist der Auflistung (56) zu sehen:

(56) 1-mal (2138, 93.3%), 2-mal (97, 4.2%), 3-mal (31, 1.4%), 4-mal (10, 0.4%), 7-mal (5, 0.2%), 5-mal (4, 0.2%), 10-mal (2, 0.1%), 9-mal (2, 0.1%), 11-mal (1, 0.0%), 8-mal (1, 0.0%)

Die lexikalische Füllung der Wort-Konjunkte in NEGRA ist in (57) aufgelistet. Neben vielen Plural-Koordinationen, welche durch die Verknüpfung von weiblichen mit männlichen Bezeichnungen erscheinen wie in „Schülerinnen Schüler“, sind es Rubriküberschriften wie „Namen [+] Notizen“, domänenspezifische Organisationsnamen, Personen- und Ortsbezeichnungen. Es treten auch einige Zwillingsformeln, auch Binominale genannt (Lambrecht 1984), mehrfach auf: „Ost [und] West“, „Art [und] Weise“, „Jahr [für] Jahr“.

Kategorie	Anzahl	in %	Wortkonjunkte	Anzahl	in %
CNP	5169	52.0	W	2577	25.9
			P	1868	18.8
			WP	724	7.3
CS	2555	25.7	P	2500	25.1
			WP	54	0.5
			W	1	0.0
CAP	900	9.1	W	733	7.4
			WP	97	1.0
			P	70	0.7
CVP	547	5.5	P	444	4.5
			WP	88	0.9
			W	15	0.2
CPP	477	4.8	P	472	4.7
			WP	5	0.1
CO	166	1.7	P	102	1.0
			WP	56	0.6
			W	8	0.1
CAVP	95	1.0	W	86	0.9
			P	6	0.1
			WP	3	0.0
CAC	26	0.3	W	24	0.2
			WP	2	0.0
CVZ	5	0.1	P	4	0.0
			WP	1	0.0
CCP	1	0.0	WP	1	0.0

Tabelle 2.28: Verhältnis aller koordinierten Phrasen bezüglich der Wortkonjunkte in NEGRA. Legende zu den Kürzeln in der Spalte „Wortkonjunkte“: W= nur Konjunkte aus Einzelwörtern. WP= mindestens ein Konjunkt aus einem Einzelwort. P= kein Konjunkt aus einem Einzelwort

(57) Lexikalische Füllungen mit mindestens 2 Vorkommen:
„Schülerinnen Schüler“ (11), „CDU FDP“ (10), „SPD Grüne“ (10), „ARD ZDF“ (9), „Kinder Jugendliche“ (9), „Städte Gemeinden“ (8), „Bürgerinnen Bürger“ (7), „Jungen Mädchen“ (7), „Kaffee Kuchen“ (7), „Männer Frauen“ (7), „Namen Notizen“ (7), „Damen Herren“ (5), „Frauen Männer“ (5), „Jens Alex“ (5), „Sinti Roma“ (5), „AFP Reuter“ (4), „Art Weise“ (4), „Gewerkschafterinnen Gewerkschafter“ (4), „Industrie- Handelskammer“ (4), „Kirchen Religionsgemeinschaften“ (4), „Mädchen Jungen“ (4), „Ost West“ (4), „Samstag Sonntag“ (4), „Seniorinnen Senioren“ (4), „Tag Nacht“ (4), „AFP dpa“ (3), „April Mai“ (3), „Arbeit Berufsausbildung“ (3), „Aus- Fortbildung“ (3), „Beherrschungs- Gewinnabführungsvertrag“ (3), „Besucherinnen Besucher“ (3), „Bornheim Ostend“ (3), „Bürger Bürgerinnen“ (3), „CDU/CSU FDP“ (3), „Dornbusch Eschersheim Ginnheim“ (3), „Düsseldorf Köln“ (3), „Erziehung Wissenschaft“ (3), „Frauen Männern“ (3), „Geborgenheit Sicherheit“ (3), „Hartmann Braun“ (3), „Haupt-Finanzausschuß“ (3), „Heimat- Geschichtsverein“ (3), „Jahr Jahr“ (3), „Kroatien Slowenien“ (3), „Kunst Leben“ (3), „Ost- Westdeutschland“ (3), „Rat Hilfe“ (3), „SPD CDU“ (3), „SPD Grünen“ (3), „Serben Kroaten“ (3), „Spielerinnen Spieler“ (3), „Stadtmitte Ostend“ (3), „Stadt Land“ (3), „Taiwan Singapur Malaysia“ (3), „Teilnehmerinnen Teilnehmer“ (3), „Zuwendung Beachtung“ (3)

Aufbau von CNP-Koordinationen Die Tabelle 2.29 auf der nächsten Seite enthält alle reinen Einwort-Konjunkte. Da es sehr viele rare Vorkommen gibt, sind aus Platzgründen alle Konjunkte, welche weniger als 3 Mal vorkommen, nicht dargestellt. Die Prozentangaben beziehen sich jedoch auf alle Vorkommen.

Die zweigliedrigen syndetischen Konstruktionen machen über 75% der insgesamt 2163 Vorkommen aus.

Funktion und Mutterkategorie von Einwort-Konjunkten Die Tabelle 2.30 auf Seite 61 macht deutlich, wie viele Wortkoordinationen maximal phrasal sind. Nämlich alle, welche eine Funktion wie SB, GR oder OA tragen. Wegen der Integration der Nominalphrasenbestandteile in die PP ist leider gerade bei der häufigsten Funktions-Kategorie NK (*noun kern*) eine Beurteilung zur Phrasaliät mit dieser Information allein nur schlecht zu leisten. Im Abschnitt 2.4.2 auf Seite 63 werden die als Nominalkerne funktionierenden Knoten genauer unterschieden. Gut 2.5% der CNP stehen isoliert als „satzwertiges“ Segment, erkennbar am „—“ in den hinteren Spalten. Es handelt sich dabei meist um Überschriften wie in (58).

(58) [$_{CNP}$ Die Hochschulen und der EG-Marmeladentopf] [N_{6654}]

Im Folgenden werden einige Bespiele für die in Tabelle 2.30[25] ausgezählten Strukturen gegeben, welche als reine Wortkoordination phrasale oder appo-

[25] Die beiden CNP in MNR-Funktion sind inkonsistent text-korrigierende Annotationen, wo die annotierende Person eine PP verstanden hat, aber trotzdem die im Text ausbuchstabierte und sichtbare

in %	Anzahl	Tochterkonstituenten	kumulativ
62.8	1359	NN KON NN	63
12.5	271	NE KON NE	75
7.8	169	NN NN KON NN	83
1.9	41	NE NE	85
1.8	39	NE NE KON NE	87
1.8	38	NN NN NN	89
1.2	25	NE KON NN	90
1.1	24	NN NN NN KON NN	91
1.0	21	NN NN	92
0.9	20	NE NE NE	93
0.6	14	CARD KON CARD	93
0.5	11	NE NE NE KON NE	94
0.5	10	NN APPR NN	94
0.5	10	NE NE NE NE KON NE	95
0.4	9	CARD CARD	95
0.3	7	NN NN NN NN KON NN	96
0.3	7	NN KON NN NN KON NN	96
0.2	4	NN KON NE	96
0.1	3	NN NN NN NN NN KON NN	96
0.1	3	NN NN NN NN NN	96
0.1	3	NN NN NN NN	96
0.1	3	NN NE	96
0.1	3	NE NE NE NE	97
0.1	3	CARD CARD CARD	97

Tabelle 2.29: Verteilung der Tochterkonstituenten von CNP-Wortkoordinationen in NEGRA mit mindestens 3 Vorkommen.

in %	Anzahl	Syntaktische Funktion	Mutterkategorie	kumulativ
46.6	1008	NK	PP	47
28.1	607	NK	NP	75
11.7	254	SB	S	86
3.2	70	OA	S	90
2.6	57	—	—	92
1.7	36	APP	NP	94
1.6	34	OA	VP	96
1.1	23	CJ	CNP	97
0.8	18	APP	PP	97
0.5	10	MO	S	98
0.4	8	PNC	MPN	98
0.3	7	PD	S	99
0.3	6	MO	VP	99
0.2	5	GR	PP	99
0.2	4	RE	NP	99
0.1	2	DA	S	99
0.1	2	CJ	CO	100
0.1	2	APP	VP	100
0.0	1	SB	VP	100
0.0	1	PD	VP	100
0.0	1	MO	AP	100
0.0	1	MNR	PP	100
0.0	1	MNR	NP	100
0.0	1	GR	NP	100
0.0	1	GL	NP	100
0.0	1	DA	VP	100
0.0	1	DA	AP	100
0.0	1	APP	S	100

Tabelle 2.30: Verteilung der Funktion der CNP-Wortkoordinationen in NEGRA.

sitive Funktion erfüllen: In (59a) ein Subjekt mit generischer Singularlesart, in (59b) ein Akkusativobjekt in einem aus stilistischen Gründen begleiterfrei gerafftem Schreibstil. In (59c) mit zwei reinen Pluralformen ein Akkusativobjekt, in (59d) ein Dativobjekt. In (59e) zu einem postnominalen Genitiv koordinierte Eigennamen, in (59f) dasselbe pränominal. In (59h) als Prädikativkonstruktion zum Subjekt. In (59i) als nominales Temporaladverbiale. In (59j) mit der Funktion PNC, wo die CNP koordinierte Vornamen zusammenfasst[26].

(59) a. [...] [$_{CNP-SB}$[$_{NN-CJ}$ Gebetskreis] , [$_{NN-CJ}$ Kinderstunde] [$_{KON-CD}$ und] [$_{NN-CJ}$ Bibelstunde]] versuchen zu zeigen, [...] [N_{2977}]

b. Anita Oster ist Angestellte, zahlt [$_{CNP-OA}$ [$_{NN-CJ}$ Sozialversicherung] [$_{KON-CD}$ und] [$_{NN-CJ}$ Lohnsteuer]] , Nettogehalt 1400 Mark. [N_{3022}]

c. [...] um [$_{CNP-OA}$[$_{NN_{CJ}}$ Transparente] [$_{KON-CD}$ oder] [$_{NN-CJ}$ Lichterketten]] aufzuhängen [...] [N_{3063}]

d. [...] seit die Änderung des Hessischen Kommunalabgabengesetzes [$_{CNP-DA}$ [$_{NN-CJ}$ Städten] [$_{KON-CD}$ und] [$_{NN-CJ}$ Gemeinden]] erlaubt, [...] [N_{13186}]

e. Wie das künftig in der Fraktion nach dem Rücktritt [$_{CNP-GR}$ [$_{NE-CJ}$ Schwappachers] , [$_{NE-CJ}$ Meinkes] [$_{KON-CD}$ und] [$_{NE-CJ}$ Weimars]] durch Nachrücker aussehen wird, [...] [N_{15826}]

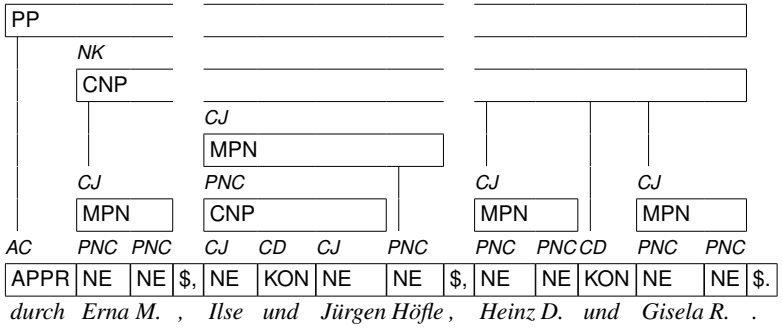
f. [...] knüpft sich nun [$_{CNP-GL}$ [$_{NE-CJ}$ Welschs] [$_{KON-CD}$ und] [$_{NE-CJ}$ Wellmers]] Interpretation von Adornos Erhabenheitsästhetik. [N_{1822}]

g. Unter Berücksichtigung [$_{NP}$ der biochemischen Verhältnisse - [$_{CNP-APP}$ [$_{NN-CJ}$ Herzzeitvolumen] [$_{KON-CD}$ und] [$_{NE-CJ}$ Druckarbeit]] - sollten [...] [N_{2215}]

h. [$_{NP-SB}$ Mexikanische Politik] ist [$_{CNP-PD}$ [$_{NN-CJ}$ Almachie] [$_{KON-CD}$ und] [$_{NN-CJ}$ Tradition]]. [N_{649}]

i. Wir haben uns [$_{CNP-MO}$ [$_{NN-CJ}$ Tage] [$_{KON-CD}$ und] [$_{NN-CJ}$ Nächte]] unterhalten. [N_{2284}]

j.



[N_{3693}]

CNP-Struktur genommen hat.

[26]Das Problem der Eigenamenbehandlung in NEGRA, welche keine Trennung zwischen Kernnamensbestandteil und enger Apposition macht, wird bei den phrasalen CNP vertieft diskutiert.

NEGRA (total 42287)		
in %	Anzahl	Kode
64.2	27149	1
34.5	14578	2
1.2	518	0
0.1	41	3

TIGER (total 86917)		
in %	Anzahl	Kode
63.4	55098	1
35.6	30935	2
0.9	801	0
0.1	82	3

Tabelle 2.31: Verteilung der NK-Tochterkonstituenten in den NP von NEGRA und TIGER. Legende: 0 = keine NK-Konstituente, 1 = nur NK-Konstituenten, 2 = Konstituenten mit anderer Funktion befinden sich vor oder nach den NK-Konstituenten, 3 = Konstituenten mit anderer Funktion befinden sich zwischen NK-Konstituenten

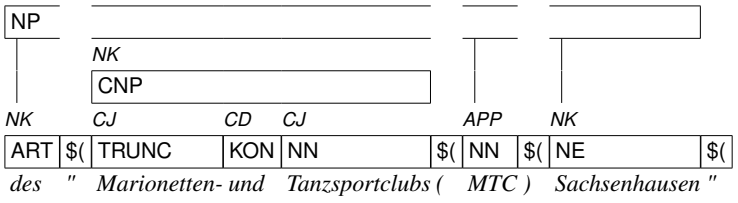
Zur Einbettung von Nominalkernen (NK) Die Funktion NK ist keine linguistisch eindeutige Abhängigkeit wie etwa Subjekt (SB) oder Akkusativobjekt (OA). Im NEGRA-Annotationsmodell wurde bewusst darauf verzichtet, bei der NP eine Entscheidung zu fällen, ob der nominale Kern oder die Artikel/Determiner-Kategorie den Kopf der NP bzw. DP bildet, ist doch in der Linguistik seit der Dissertation von Abney (1987) die DP-Analyse immer wichtiger geworden ist. Nominale Kerne (inklusive enger Appositionen) erhalten ebenso wie deren Begleiter und attributive Adjektive die Funktion NK annotiert. Diese Form von Theorieneutralität erweist sich für automatische Auswertungen, welche Köpfe zuverlässig erkennen will, als mühsam.

Um die folgende Datenzusammenstellung richtig interpretieren zu können, ist es wichtig, dass trotz der wenig spezifischen Auffassung darüber, was ein NK darstellt, die Topologie der grammatischen Funktionen bei NEGRA in der NP überblickt wird.

In der Tabellenzusammenstellung 2.31 sieht man, dass die Tochterkonstituenten in NK-Funktion bis auf seltenste Ausnahmen innerhalb einer NP an einem Stück erscheinen, d.h. nicht durch andere Funktionen unterbrochen werden.

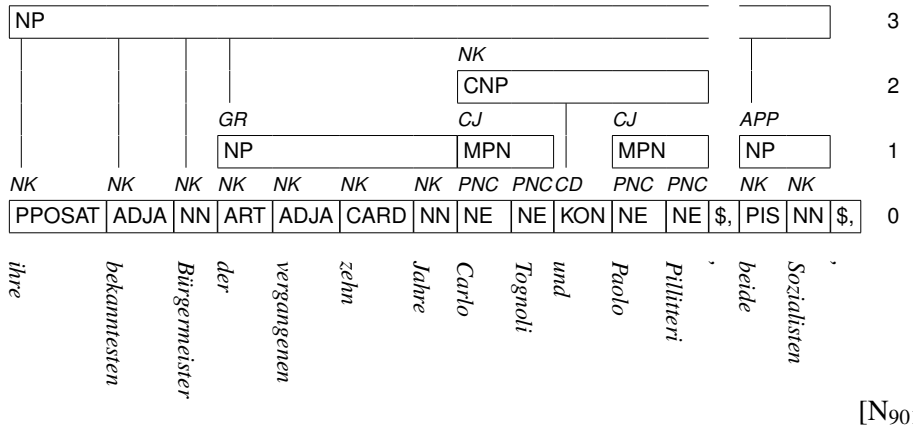
Die Fälle mit unterbrochener NK-Sequenz sind insbesondere Abkürzungen und Akronyme in Klammern, welche als Apposition (APP) annotiert werden, die darauf nachfolgende enge Apposition hingegen wieder als NK wie in Beispiel (60a). Oder enge Appositionen, welche rechts von postnominalen Genitiven (GR) erscheinen, und ebenfalls als NK annotiert sind wie in (60b).

(60) a.



[N_{3430}]

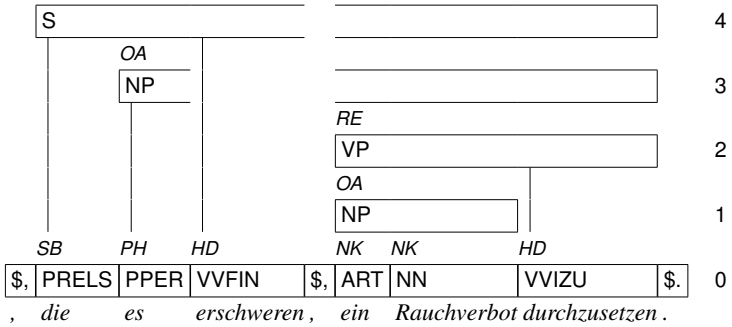
b.



[N_{9019}]

Bei den gut 500 Fällen, wo gar kein Nominalkern annotiert ist (Kode 0 in Tabelle 2.31 auf der vorherigen Seite), handelt es sich insbesondere um Konstruktionen mit expletivem „es“ wie in (61a) und weitere als Platzhalterkonstruktionen[27] aufgefasste Phänomene wie in (61b).

(61) a.



[N_{10704}]

b. [$_{NP}$ [$_{NN-RE}$ Expansion] , [$_{PDS-PH}$ das]] ist kein Thema für uns. [N_{3195}]

Wenn man quantitativ überblicken will, welche Rolle eine CNP innerhalb einer NP oder PP spielt, muss der Kontext betrachtet werden. Dies erlaubt, die notwendigen Rückschlüsse auf die Struktur der einbettenden PP oder NP zu machen.

Die Einbettung der CNP in NP Die Tabellenzusammenstellung 2.32 auf der nächsten Seite zeigt die unmittelbare linke bzw. rechte Nachbarschaft in Form ihrer Funktion, d.h. die Konfiguration:

$$[_{NP} \ldots [_{?-\textbf{linkeSF}} \ldots] [_{CNP} \ldots] [_{?-\textbf{rechteSF}} \ldots] \ldots]$$

In knapp 70% der Fälle geht der CNP mindestens ein weiteres NK-Element voraus. In knapp 20% wird nur die rechte Seite der CNP ergänzt. Pränominale

[27] Stichprobenartiges Überprüfen solcher Konstruktionen zeigen meines Erachtens einige Schwierigkeiten in der Annotation auf. Das NEGRA-Annotationshandbuch (Brants u. a. 1999, 62) enthält dazu auch nur wenig Informationen.

in %	Anzahl	linke SF
69.5	422	NK
18.8	114	—
6.6	40	MO
4.1	25	CM
0.7	4	MNR
0.2	1	NG
0.2	1	GL

in %	Anzahl	rechte SF
56.7	344	—
14.2	86	MNR
13.2	80	GR
5.8	35	RC
4.6	28	NK
4.4	27	APP
1.0	6	PG
0.2	1	MO

Tabelle 2.32: Verteilungen der Funktionen der linken und rechten Schwesterkonstituenten von CNP-Wortkoordinationen innerhalb von total 607 NP.
Legende: SF = Schwesterfunktion

Modifikatoren (MO) bestehen entweder aus Adverbien oder PP. Konjunktionalphrasen, bestehend aus Vergleichskonjunktionen (CM wie „als“ oder „wie“) und den Elementen einer Nominalphrase, werden in NEGRA analog nicht-einbettend zur PP strukturiert und ergeben immer eine NP-CC wie in (62).

(62) Rolf Müller, zuständiger Redakteur für HR-3-Sendungen [$_{NP-CC}$ [$_{KOKOM-CM}$ wie] [$_{CNP}$ " [$_{NN}$ Extra] " [$_{KON}$ und] " [$_{NN}$ Graffity]]]", hat andere Konzepte im Kopf. [N_{294}]

Die kategorielle Füllung der total 494 nicht-leeren linken Schwestern ist in (63) aufgelistet. Erwartungsgemäss sind Artikel und Adjektive dominant.

(63) Verteilung der Kategorien der linken Schwestern von CNP-Wortkoordinationen innerhalb von NP mit mehr als 1% relativer Häufigkeit:
ART (31.2%), ADJA (19.1%), NN (16.8%), ADV (6.1%), AP (5.5%), KOKOM (5.1%), PIDAT (3.2%), PPOSAT (2.8%), CARD (2.8%), PIAT (2.0%), PP (1.8%)

Keine Ergänzung auf der rechten Seite ist in über der Hälfte der CNP der Fall und das Auftreten von NK mit knapp 5% an dieser Stelle wenig häufig.

Wenn man die linken und rechten Kontexte nicht isoliert betrachtet, zeigt sich, dass die CNP mehrheitlich an der rechten Phrasengrenze der Mutter steht, insgesamt aber viele verschiedene Kombinationen vorliegen. Die Tabelle 2.33 auf der nächsten Seite weist dies für alle Kontexte mit mindestens 3 Vorkommen nach.

Die Einbettung der CNP in PP Die Tabelle 2.34 auf der nächsten Seite zeigt die Funktion der unmittelbaren linken und rechten Nachbarschaft innerhalb von PP. In knapp 75% der Fälle, nämlich dort, wo die linke Schwester die Präposition ist (erkennbar an der Funktion AC) und die rechte Schwester nicht existiert ('—'), liegt eine echte Wortkoordination vor. Die restlichen CNP werden auf der Ebene der eingebetteten Nominalphrase noch erweitert.

in %	Anzahl	linke SF	rechte SF	kumulativ
47.0	285	NK	—	47
7.6	46	—	MNR	55
7.2	44	NK	GR	62
5.6	34	NK	MNR	67
5.6	34	MO	—	73
5.6	34	—	GR	79
3.6	22	NK	RC	82
3.5	21	CM	—	86
3.3	20	NK	NK	89
2.1	13	—	RC	91
2.1	13	—	APP	93
2.0	12	NK	APP	95
1.0	6	—	NK	96
0.7	4	NK	PG	97
0.5	3	CM	MNR	97

Tabelle 2.33: Verteilung der Funktionen der Schwesterkonstituenten von CNP-Wortkoordinationen innerhalb von total 607 NP in NEGRA.
Legende: SF = Schwesterfunktion

in %	Anzahl	linke SF	rechte SF	kumulativ
74.2	748	AC	—	74
13.8	139	NK	—	88
2.7	27	AC	GR	91
2.6	26	AC	MNR	93
1.4	14	NK	GR	95
1.2	12	AC	RC	96
1.1	11	NK	MNR	97
0.6	6	AC	APP	98
0.5	5	NK	RC	98
0.4	4	NK	PG	98
0.4	4	NK	NK	99

Tabelle 2.34: Verteilung der Funktionen der Schwesterkonstituenten von CNP-Wortkoordinationen innerhalb von total 1008 PP in NEGRA. Angezeigt werden Typen mit mindestens 3 Vorkommen.
Legende: SF = Schwesterfunktion

2.4.3 CAP

Wie in Tabelle 2.43 auf Seite 101 ersichtlich, bilden CAP- Wortkoordinationen in NEGRA mit 733 Vorkommen den zweitgrössten Anteil aller Wortkoordinationen. Der Anteil der reinen Wortkoordinationen ist folgendermassen verteilt: Wortkoordination (735, 81.7%), Wortgruppenkoordination (165, 18.3%).

Wenn wir nur die lexikalischen Konjunkte betrachten und die Konjunktoren ausser Acht lassen, ergeben sich insgesamt 647 verschiedene Types mit einem fast ausgeglichenen Type-Token-Verhältnis von 1:1.1 . Die exakte Verteilung der Vorkommenshäufigkeit ist in der Auflistung (64) zu sehen:

(64) 1-mal (599, 92.6%), 2-mal (31, 4.8%), 3-mal (10, 1.5%), 4-mal (2, 0.3%), 8-mal (2, 0.3%), 5-mal (1, 0.2%), 6-mal (1, 0.2%), 7-mal (1, 0.2%)

Die lexikalische Füllung der Wort-Konjunkte in NEGRA ist in (65) aufgelistet. Neben vielen Zahlausdrücken finden sich einige Zwillingsformeln wie „frank [und] frei“ oder „klipp [und] klar“.

(65) Lexikalische Füllungen mit mindestens 2 Vorkommen in NEGRA:
„13. 14.“ (8), „zwei drei“ (8), „vier fünf“ (7), „Berger Bischofsheimer“ (6), „jung alt“ (5), „14 18“ (4), „drei vier“ (4), „10 15“ (3), „20 30“ (3), „30 40“ (3), „60 70“ (3), „70 80“ (3), „acht zehn“ (3), „direkt indirekt“ (3), „drei fünf“ (3), „drei sechs“ (3), „ein zwei“ (3), „10 12“ (2), „14 16“ (2), „15 10“ (2), „16. 17.“ (2), „200 300“ (2), „20 60“ (2), „45 60“ (2), „4 5“ (2), „60 100“ (2), „60 90“ (2), „7 10“ (2), „dritten vierten“ (2), „einer zwei“ (2), „eine andere“ (2), „frank frei“ (2), „fünf sechs“ (2), „fünf zehn“ (2), „kleinen großen“ (2), „klipp klar“ (2), „kroatische moslemische“ (2), „mehr bessere“ (2), „perspektivlos entmutigend“ (2), „sechs acht“ (2), „sechs zehn“ (2), „sechs zwölf“ (2), „serbischer kroatisch-moslemischer“ (2), „sieben zehn“ (2), „siebzehn zwanzig“ (2), „zehn zwölf“ (2), „zweite dritte“ (2), „zwei vier“ (2)

Bau der CAP-Wortkoordinationen Die 735 Wortkoordinationen wiederum bestehen fast ausschliesslich wie in Beispiel (67a) aus 3 Elementen (2 Konjunkten und 1 Konjunktor), wie die folgende Auflistung zur Anzahl der Konjunktglieder zeigt:

(66) 3-gliedrig (669, 91.0%), 5-gliedrig (42, 5.7%), 4-gliedrig (11, 1.5%), 7-gliedrig (6, 0.8%), 2-gliedrig (4, 0.5%), 9-gliedrig (1, 0.1%), 6-gliedrig (1, 0.1%), 11-gliedrig (1, 0.1%)

Wenn man die Wortarten der Konjunktglieder betrachtet, wie sie in Tabelle 2.35 auf der nächsten Seite zusammengestellt sind, ergibt sich ein sehr einheitliches Bild: attributive Adjektive (ADJA), Kardinalzahlen (CARD) sowie prädikative oder adverbiale Adjektive (ADJD) bilden homogene Koordinationsphrasen.

in %	Anzahl	Konjunktkategorien	kumulativ
36.3	267	ADJA ADJA	36
32.9	242	CARD CARD	69
19.2	141	ADJD ADJD	88
3.3	24	ADJA ADJA ADJA	92
3.0	22	TRUNC ADJA	95
1.2	9	ADJD ADJD ADJD	96
1.0	7	TRUNC ADJD	97
0.7	5	ADJA ADJA ADJA ADJA	98
0.5	4	CARD CARD CARD	98
0.3	2	ADJD	98
0.3	2	PIAT ADJA	99
0.1	1	ADJA ADJA ADJA ADJA ADJA	99
0.1	1	ADJA ADJA ADJA ADJA ADJA ADJA	99
0.1	1	ADJA ADJD ADJD	99
0.1	1	ADJD ADJD ADJD ADJD	99
0.1	1	ADJD VVPP	99
0.1	1	ART ADJA	99
0.1	1	ART ART	99
0.1	1	ART PIAT	99
0.1	1	PIS ADJA	100
0.1	1	TRUNC TRUNC ADJA	100

Tabelle 2.35: Verteilung der Tochterkonjunkte der CAP-Wortkoordinationen in NEGRA

Eine Quelle von Unsicherheit und (Annotations-)Fehlern ist mit dem Lemma „ein“ verbunden. Im Beispiel (67a) liegt die unflektierte Verwendung von „ein“ in der Fügung „ein oder ander“[28] vor, welche man mit „mancher“ paraphrasieren kann. Gemäss STTS-Handbuch (Schiller u. a. 1999, 33) ist „eine“ als ADJA zu bestimmen in Wendungen wie „der eine/ADJA oder andere/ADJA Mensch“.

Etwas anders liegt der Fall in Beispiel (67b) mit dem flektierten „eine“. Die Analyse von (Dipper 2003, 158) nimmt sowohl „eine“ wie „mehrere“ aus morphologischen Gründen als quantifizierende Adjektive (AQUANT) an, wobei für „eine“ auch noch eine Begleiterfunktion existiert.

(67) a. [...] und der [$_{CAP}$[$_{ART-CJ}$ ein] oder [$_{ADJA}$andere]] Sportwagen dreht vor giftgrünen Hügeln seine Runden . [N$_{10338}$]

b. Das erste Stadium kann [$_{CAP}$[$_{ART-CJ}$ eine] oder [$_{PIAT}$ mehrere]] Stunden andauern . [N$_{8023}$]

Koordinierte Begleiter Wie in Tabelle 2.35 auf der vorherigen Seite ersichtlich, gibt es auch eine CAP, welche nur aus koordinierten definiten Begleitern besteht wie in Beispiel (68):

(68) [$_{CAP}$ [$_{ART}$ Der oder [$_{ART}$die] " BerichterstatterIn " mag sich gedacht haben [...] [N$_{6482}$]

Solche Konstruktionen sind im Zusammenhang mit Bemühungen um nichtdiskriminierenden Sprachgebrauch typisch, es fehlt in NEGRA (und auch in TIGER) jedoch eine Kategorie für koordinierte Begleiter. In TIGER werden koordinierte Begleiter als Verlegenheitslösung mit CO annotiert wie in Beispiel (69):

(69) Allerdings werden [$_{NP}$[$_{CO-NK}$[$_{ART}$ der] oder [$_{ART}$ die]] neuen Anbieter] kaum vor dem Sommer 1997 loslegen dürfen. [T$_{22948}$]

2.4.4 CAVP

Da in NEGRA gemäss Tabelle 2.1 auf Seite 13 nur 95 koordinierte Adverbien vorkommen, die etwa 1% aller Vorkommen von koordinierten Strukturen ausmachen, wird im Folgenden über allen 3 Korpora ausgewertet. Insgesamt liegen 281 CAVP vor, wobei sich der Anteil der reinen Wortkoordinationen folgendermassen verteilt: Wortkoordination (252, 89.7%), Wortgruppenkoordination (29, 10.3%) .

Die 252 Wortkoordinationen wiederum bestehen fast ausschliesslich wie in Beispiel (70a) aus 3 Elementen (2 Konjunkten und 1 Konjunktor), wie die folgende Auswertung zur Anzahl der Konjunktglieder zeigt: 3-gliedrig (246, 97.6%), 5-gliedrig (3, 1.2%), 2-gliedrig (2, 0.8%), 4-gliedrig (1, 0.4%) .

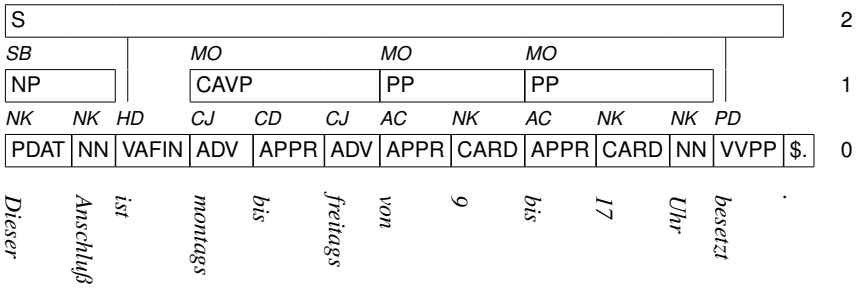
Neben den gleichordnenden Konjunktoren ist „wie“ wie in Beispiel (70b) sowie „bis“ wie in Beispiel (70c) gängig. Letzteres allerdings wird in NEGRA nur als Konjunktor annotiert, wenn kein „von“ vorausgeht.

[28] Gemäss Artikel zu „ein“ in Klosa und Auberle (2001) ist diese unflektierte Form noch bei „ein oder/und/bis zwei“ üblich.

(70) a. Weil derartige Überlegungen aber einer freudigen Stimmung [$_{CAVP-MO}$ [$_{ADV}$ ganz] [$_{KON}$ und] [$_{ADV}$ gar]] nicht zuträglich sind, [...] [N$_{782}$]

b. Es war [$_{CAVP-MO}$ [$_{ADV}$ damals] [$_{KON}$wie] [$_{ADV}$heute]] auf drei Seiten vom Friedhof umgeben. [N$_{4387}$]

c.



[N$_{1406}$]

Typisch für CAVP auf Wortebene sind feste Wendungen, welche einen starken Mehrwortlexem-Charakter aufweisen. [29]

Nachfolgend sind alle Wortfüllungen von CAVP in Kleinschreibung aufgeführt mit ihrer Häufigkeit in Klammern:

(71) „nach wie vor“ (66), „mehr oder weniger“ (12), „nach und nach“ (12), „montags bis freitags“ (8), „hin und her“ (7), „so oder so“ (5), „mehr und mehr“ (4), „wohl oder übel“ (4), „ab und an“ (3), „auch und gerade“ (3), „da und dort“ (3), „dienstags bis freitags“ (3), „durch und durch“ (3), „hier und da“ (3), „hin und wieder“ (3), „kreuz und quer“ (3), „morgens und abends“ (3), „ab und zu“ (2), „auf und ab“ (2), „damals und heute“ (2), „früher oder später“ (2), „ganz und gar“ (2), „hier wie dort“ (2), „hie und da“ (2), „links und rechts“ (2), „mehr oder minder“ (2), „oben und unten“ (2), „ost und west“ (2), „rechts oder links“ (2), „sehr , sehr“ (2), „treppauf - treppab“ (2), „wann und wo“ (2), „wo und wie“ (2), „50- bis 100mal“ (1), „angst und bange“ (1), „auf und davon“ (1), „außen und innen“ (1), „bestens und international“ (1), „damals wie heute“ (1), „dann und dann“ (1), „dann und wann“ (1), „draußen und drinnen“ (1), „drei- bis viermal“ (1), „drei- oder viermal“ (1), „drei- und fünfmal“ (1), „drei bis viermal“ (1), „drunter und drüber“ (1), „eh und je“ (1), „eigentlich eigentlich“ (1), „ein- bis zweimal“ (1), „einst und heute“ (1), „ein für allemal“ (1), „ein noch aus“ (1), „frieher un heut“ (1), „für und wider“ (1), „ganz , ganz“ (1), „ganz oder teilweise“ (1), „gerade und zuerst“ (1), „gern und häufig“ (1), „gern und oft“ (1), „heute und hier“ (1), „heute und morgen“ (1), „hierzulande und draußen“ (1), „hier und dort“ (1), „hier und jetzt“ (1), „hinten

[29](Steiner 2003, 109) enthält im Anhang eine Liste von Mehrwortlexemen im Münsteraner Korpus, welche auch einige Adverbkoordinationen auflistet: „nach wie vor“ (16), „durch und durch“ (7), „ganz und gar“(5), „hier und da“ (5), „ab und an“ (4), „ab und zu“ (4), „hin und wieder“ (3), „nach und nach“ (3),„hie und da“(2), „hin und her“ (2). Diese Mehrwortlexeme findet man fast alle in den CAVP-Füllungen, die in der Beispielsammlung (71) auf dieser Seite aufgeführt sind.

und vorne“ (1), „innen wie außen“ (1), „irgendwann , irgendwie“ (1), „ja oder nein“ (1), „landauf landab“ (1), „lange und weit“ (1), „mal für mal“ (1), „miteinander statt gegeneinander“ (1), „mittwochs , donnerstags und freitags“ (1), „montags , dienstags und donnerstags“ (1), „montags bis donnerstags“ (1), „montags bis mittwochs“ (1), „montags und donnerstags“ (1), „morgens und mittags“ (1), „nicht - oder allzugut“ (1), „nicht oder zuwenig“ (1), „oft und gern“ (1), „oft und reichlich“ (1), „plus / minus“ (1), „politisch und militärisch“ (1), „rauf und runter“ (1), „rechts wie links“ (1), „samstags und sonntags“ (1), „selten oder nie“ (1), „so oder ähnlich“ (1), „toi , toi , toi“ (1), „überall und nirgends“ (1), „über und über“ (1), „unten oder oben“ (1), „vornehmlich und berechtigterweise“ (1), „vorn und hinten“ (1), „vorwärts oder rückwärts“ (1), „vorwärts und rückwärts“ (1), „wann und wie“ (1), „warum und wie“ (1), „wech oder dazu“ (1), „wieder und wieder“ (1), „wie und wann“ (1), „wie und warum“ (1), „wie und wo“ (1), „zusammen und durcheinander“ (1), „zwei- , dreimal“ (1), „zwei- bis dreimal“ (1)

Der Ausdruck „nach wie vor“ macht beinahe 1/4 aller Vorkommen aus. Es gibt zudem relativ viele Kombinationen mit Wochentagsadverbien, was insbesondere mit den Veranstaltungshinweisen in diesem Zeitungstextkorpus zusammenhängt. Auffällig sind weiter die beiden asyndetischen „sehr, sehr“ und „ganz, ganz“, welche in der Umgangssprache häufig[30] zur Intensivierung verwendet werden. Die Schreibung mit Komma scheint im Deutschen üblich zu sein. Im ausgewogenen „DWDS Kerncorpus Version 170605b“[31] findet sich die Schreibung ohne Komma nur 27 Mal, mit Komma 327 Mal.

(72) a. die Melodien sind [$_{AP}$ [$_{CAVP-MO}$ [$_{ADV-CJ}$ sehr] , [$_{ADV-CJ}$ sehr]] [$_{ADJD-HD}$ eingängig]] . [N_{23}]

b. Dabei gäbe es gute Gründe, sie [$_{AP}$ [$_{CAVP-MO}$ [$_{ADV-CJ}$ ganz] , [$_{ADV-CJ}$ ganz]] [$_{ADJD-HD}$ schnell]] hervorzukramen. [T_{37178}]

Ähnlich reduplizierend, aber fest gefügt und syndetisch immer mit dem Konjunktor „und“ oder „für“ verknüpft, sind die Fälle wie in „dann und dann“, „durch und durch“, „wieder und wieder“ oder „mal für mal“, bei denen die spezifische Gesamtbedeutung sich erst in der Koordination ergibt.

Bei „Ost und West“ aus [$T_{7536,7549}$] (im Original gross geschrieben) wurde eine typisch nominale Paarformel Müller (1997) als CAVP annotiert, was angesichts der durchgängigen Grossschreibung, in der sie 666 Mal im „DWDS Kerncorpus“ belegt ist, als Fehlannotation zu werten ist.

Die Verbindung „treppauf, treppab“ hingegen ist wie in Beispiel (73) eine adverbiale Paarformel, welche im Text zwischen Anführungszeichen als Name eines

[30] Eine Web-Anfrage zu diesen Formen bringt Hundertausende von Belegen. Auch im Französischen mit „très très“ oder Englischen mit „very very“ sind analoge Konstruktionen häufig.

[31] Dieses Korpus ist online abfragbar unter `http://www.dwds.de`. Hintergrundinformation zum Projekt findet sich in (Geyken 2004).

NEGRA			NEGRA, TIGER, CZ		
in %	Anzahl	Konjunkte	in %	Anzahl	Konjunkte
81.4	70	ADV ADV	84.1	212	ADV ADV
5.8	5	PWAV PWAV	3.6	9	PWAV PWAV
2.3	2	ADV ADJD	2.8	7	TRUNC ADV
2.3	2	ADV ADV ADV	1.6	4	APPR APPR
2.3	2	APPR APPR	1.6	4	PTKVZ PTKVZ
2.3	2	PTKVZ PTKVZ	1.2	3	ADV ADJD
1.2	1	ADJD ADV	1.2	3	ADV ADV ADV
1.2	1	ART ADV	0.8	2	ADV APPR
1.2	1	TRUNC ADV	0.4	1	ADJD ADJD
			0.4	1	ADJD ADV
			0.4	1	ADV PROAV
			0.4	1	ART ADV
			0.4	1	CARD ADV
			0.4	1	PROAV PROAV
			0.4	1	PTKNEG ADJD
			0.4	1	PTKNEG ADV

Tabelle 2.36: Verteilung der Tochterkonjunkte in CAVP- Wortkoordinationen in NEGRA, TIGER und CZ

Fotowettbewerbs verwendet wird.

(73) Wie zu jedem anderen Thema können auch zu " [$_{CAVP-MO}$ [$_{ADV-CJ}$ Trepp-auf] - [$_{ADV-CJ}$ treppab]] " maximal drei Beiträge eingeschickt werden. [N$_{19658}$]

Wie in Tabelle 2.36 ersichtlich, treten als Konjunkte neben Adverbien (ADV) im engen Sinn auch noch adverbiale Interrogativ- bzw. Relativpronomen (PWAV), Pronomialadverbien (PROAV), aber auch nicht-attributive Adjektive (ADJD) auf. Die Fälle mit unterschiedlichen Konjunkttypen müssten (obwohl selten sinnvoll) streng nach NEGRA-Annotationskonvention als nicht-symmetrische Koordination (CO) annotiert werden.

2.4.5 CAC

Koordinierte Präpositionen treten selten auf. Von den knapp 36'000 Vorkommen von Präpositionen im NEGRA-Korpus (28'984 davon APPR, 6'710 davon APPR-ART, 106 davon APPO) sind nur etwa 50 koordiniert, im Schnitt taucht dort etwa alle 800 Satzeinheiten eine solche Konstruktion auf.

In der Tabelle 2.37 auf der nächsten Seite sind alle Knoten der Kategorie CAC aus NEGRA zusammengestellt, welche als komplexe Präposition innerhalb einer

in %	Anzahl	Tochterkonstituenten
68.0	17	APPR KON APPR
20.0	5	APPRART KON APPRART
8.0	2	APPR KON AVP
4.0	1	APPRART KON APPR

Tabelle 2.37: Verteilung der Tochterkonstituenten aller 25 Fälle von koordinierten Präpositionen in NEGRA

Präpositionalphrase fungieren. Hier verhält sich die Anzahl der koordinierten APPR und APPRART ungefähr proportional zur Anzahl der nicht-koordinierten Vorkommen. Wegen der kleinen Datenmenge ist es trotzdem statistisch nicht aussagekräftig – tatsächlich sehen die Verhältnisse im TIGER-Korpus ganz anders aus. Dort findet sich keine einzige APPRART-Koordination.

In der Auflistung (74) sind alle in CAC auftretenden Präpositionen zusammengestellt. Die 14 Types ergeben ein Type-Token-Verhältnis von 1:1.7. Nur die Kombination „in um“ sticht heraus, welche in den meisten Fällen dem Muster „in und um STADTNAME“ zu verdanken ist.

(74) Lexikalische Füllungen in NEGRA:
„in um“ (6), „im am“ (2), „mit ohne“ (2), „von nach“ (2), „vor nach“ (2), „vor während“ (2), „am im“ (1), „im um“ (1), „im ums“ (1), „in an“ (1), „nach nach“ (1), „nach von“ (1), „vom ums“ (1), „während nach“ (1)

Wenn man sich die lexikalischen Füllungen aller 20 CAC-Koordinationen aus TIGER in der Auflistung (75) anschaut, erkennt man nur wenig Übereinstimmung mit NEGRA.

(75) Lexikalische Füllungen in TIGER:
„mit ohne“ (3), „innerhalb außerhalb“ (2), „in um“ (2), „ab bis“ (1), „an in“ (1), „dies- jenseits“ (1), „diesseits jenseits“ (1), „für gegen“ (1), „in aus“ (1), „in von“ (1), „in vor“ (1), „mit für“ (1), „unter über“ (1), „von nach“ (1), „von zur“ (1), „vor neben hinter“ (1)

Typischerweise werden diejenigen Präpositionen koordinativ verknüpft, welche die gleiche Kasusforderung stellen: Dies kann ohne Artikel wie in (76a) oder mit separatem Artikel wie in (76b) geschehen. Auch die Koordination von Präpositionen mit verschmolzenem Artikel wie in (76c) kommt vor. Diese Konstruktion wirft auf dem Hintergrund der traditionellen Vorstellung vom Aufbau von Präpositionalphrasen bestehend aus Präposition und abhängiger Nominalphrase die Frage auf, welche Funktion der Artikel im ersten Konjunkt wahrnimmt. Insbesondere, wenn man noch Beispiele wie in (77d) betrachtet, wo der Artikel einmal verschmolzen und einmal getrennt auftritt.

(76) a. [...] [$_{PP}$ [$_{CAC-AC}$[$_{APPR}$ in] [$_{KON}$ und] [$_{APPR}$ um]] [$_{NE}$ Hammersbach]]. [N_{3409}]

b. [...] [$_{PP}$ [$_{CAC-AC}$[$_{APPR}$ vor] [$_{KON}$ oder] [$_{APPR}$ nach]] [$_{ART}$ dem] [$_{NN}$ Einkauf]] . [N_{3409}]

c. [...] [$_{PP}$ [$_{CAC}$[$_{APPRART}$ im] [$_{KON}$ und] [$_{APPRART}$ am]] Gerätehaus der Stadtteilwehr [...]] [...] [N_{11067}]

Koordinierte Präpositionen mit unterschiedlichen Kasusforderungen sind keinesfalls selten: Dies kann wie in (77a) morphologisch unsichtbar realisiert werden oder eindeutig markiert erscheinen wie in (77b). Entscheidend ist die Kasusforderung der am weitesten rechts stehenden Präposition, wie es auch in normativen Grammatiken verlangt wird (Dudenredaktion 2005, §1424).

Dass es sich nicht um eine symmetrische Koordination handelt, sieht man leicht in (77c) und (77e), wenn der Konjunktor mit dem zweiten Konjunkt weggelassen wird. Der letztere Fall, wo APPR mit APPRART koordiniert ist, wird nur in wenigen normativen Grammatiken diskutiert.

(77) a. [...] das Neueste [$_{PP}$ [$_{CAC-AC}$[$_{APPRART}$ vom] [$_{KON}$ und] [$_{APPRART}$ ums]] [$_{NN}$ Fahrrad]] zu sehen [...] [N_{9185}]

b. [...] [$_{PP}$[$_{CAC}$ [$_{APPR}$ vor] [$_{KON}$ und] [$_{APPR}$ während]] [$_{ART}$ des] [$_{NN}$ Konzerts]] [...] [N_{2933}]

c. * vor des Konzerts

d. [...] tummelten sich etwa 500 Besucher [$_{PP}$ [$_{CAC}$[$_{APPRART}$ im] [$_{KON}$ und] [$_{APPR}$ um]] [$_{ART}$ das] [$_{NN}$ Zelt]] . [N_{12987}]

e. * tummelten sich etwa 500 Besucher im das Zelt

f. Derweil konnten sich die Eltern [$_{PP}$ [$_{CAC}$[$_{APPRART}$ im] [$_{KON}$ und] [$_{APPR}$ um]] [$_{ART}$ das] [$_{ADJA}$ kleine][$_{NN}$ Bierzelt] [$_{APZR}$ herum] bei Speis' und Trank die Zeit vertreiben. [N_{11109}]

Koordination von APPR und APPRART Die Wortkombination „im und um" verzeichnete am 7.7.2005 bei der Suchmaschine Google etwa 48000 Treffer auf deutschsprachigen Seiten. Es besteht somit ein Bedarf nach der kombinierten Semantik dieser beiden Präpositionen bei den Schreibenden, welche durch eine einzelne bestehende Präposition nicht ausgedrückt werden kann. Meines Erachtens besitzt diese Wortkombination starken Mehrwortlexem-Charakter und wird schon fast formelhaft als Zwillingspaar verwendet (unterstützt durch die Parallelen im silbischen Aufbau: Vokal gefolgt vom Konsonanten „m"). So ist es wenig erstaunlich, dass die semantisch im Prinzip gleichwertige Wortkombination „um und im"[32] nur einen Zehntel an Google-Treffern aufweist. Davon ist aber höchstens die Hälfte als CAC zu betrachten, wie eine stichprobenartige Auszählung ergeben hat, denn das Wort „um" ist in dieser Reihenfolge oft ein abgetrenntes Verbpräfix. Zudem gelten mehr als 800 der gefundenen Treffer der grösseren Wortkombination „rund um und im".

[32]Zu beachten ist, dass Google allfällige Interpunktionszeichen innerhalb der Wortkombination ignoriert.

Zirkumpositionen und postnominale Adverbien Das Beispiel (77f) weist noch auf die Schwierigkeiten von annotierten Zirkumpositionen bei koordinierten Präpositionen hin. Im NEGRA-Korpus wird „um ... herum“ konsequent als Zirkumposition mit „herum“ als rechter Teil annotiert. Dies steht im Widerspruch zur STTS-Annotationskonvention (Schiller u. a. 1999, 69), welche für diese Fälle ausdrücklich ADV verlangt. Auch für das Münsteraner Tagset (Steiner 2003, 48) gelten die gleichen Konventionen, wobei dort folgende distributionellen Kriterien expliziert werden:

1. Ein Adverb liegt vor, falls es unter Bedeutungserhaltung weggelassen werden kann.

2. Ein Adverb liegt vor, falls es unter Bedeutungserhaltung vor die Präposition gestellt werden kann.

Auf Grund dieser Kriterien wird dann das Wort „her“, das traditionell nur als Adverb aufgefasst wird, im Satz (78a) als Zweitglied einer Zirkumposition betrachtet – im Gegensatz zum Wort „hin“, welches wegen obiger Kriterien als Adverb kategorisiert ist. Die Annotation in NEGRA ist – wie Satz (85) belegt – in diesem Punkt ebenfalls nicht konform zum STTS, aber in sich konsistent.

(78) a. Die anderen, die aus den Schwarzwaldtälern, die stumm gegen die Freiburger Bourgeoisie des alteingesessenen Vereins ein paar hundert Meter weiter [$_{APPR}$ zur] Stadtmitte [$_{ADV}$ hin] anstehen, harren [$_{APPR}$ von] Natur [$_{APRZ}$ her] aus.
(Zitiert nach Steiner (2003, 57), ursprünglich aus: Sportclub Freiburg. Der Aufstieg der Underdogs. Zeit Nr.4, 1992)

b. Das [$_{APPR}$ vom] Repertoir [$_{APRZ}$ her] unerschrockene Ensemble möchte nur keine Langweile . [N_{59}]

Echte Zirkumposition in der GDS Die Frage, was echte Zirkumpositionen ausmacht, wird in der systematischen Grammatik (GDS) von (Zifonun u. a. 1997, 2085ff.) ebenfalls diskutiert. Als einzige eindeutige Zirkumposition akzeptiert Zifonun u. a. (1997, 47) „um ... willen“, wobei die Kriterien, welche zu dieser engen Sichtweise führen, nicht in einer Fachbegriffsbestimmung explizit gemacht werden. Das webbasierte Zusatzprojekt „Grammis“ enthält das Glossar „Das Terminologische Wörterbuch“ (IDS 2006), welches den Gehalt der GDS kompakt wiedergibt als:

> „Als Zirkumposition bezeichnet man eine im Deutschen eindeutig nur mit dem Element *um...willen* vertretene periphere Subklasse von Präpositionen. Andere Kombinationen von voran- und nachgestelltem präpositionalem Element sind keine eindeutigen Zirkumpositionen, da die Verbindung nicht fest und ein Element austauschbar ist, andere Akzentverhältnisse vorliegen oder die Stellung verändert werden kann.“

Illustrierend werden dazu noch die Beispiele in (79) gegeben.

(79) a. vom Bahnhof an/ab/aus

b. wir krochen unter/zwischen den Bäumen durch

c. unter/zwischen den Bäumen sind wir durchgekrochen

Die Wortkombination „um ... herum“ wird in (Zifonun u. a. 1997) gar nicht angesprochen, da es sich gemäss GDS bei „herum“ nicht um ein präpositionales Element handeln kann.

Die Forderung nach fester Verbindung, d.h. dass keine paradigmatische Variation im nachgestellten Element wie in (79a) oder im vorangestellten Element wie in (79b) möglich sein darf, schliesst für sie viele Kombinationen von echten präpositionalen Elementen als Zirkumposition aus. Dass man nur stark lexikalisierte Formen als Zirkumposition zulassen will, hat durchaus etwas für sich. Allerdings sollte man sich auch die Frage stellen, ob es den Status von „um..willen“ ändern würde, wenn im Sprachsystem zufälligerweise noch eine mehr oder weniger synonyme Wortkombination existierte, bei der ein Bestandteil unterschiedlich wäre.[33]

Die Veränderbarkeit in der Stellung wird von (79c) illustriert, welche die Wechselbeziehung von abtrennbaren Verbpräfixen bei Bewegungsverben mit nachgestellten präpositionalen Elementen betrifft. In ihrer inkorporierten Form tragen diese jeweils den Wortakzent („durchgekrochen“). Allerdings wird schon in Zifonun u. a. (1997, 2086) erwähnt, daß beide Varianten auch verbadjazent koexistieren wie in (80a) bzw. (80b). Insofern sind die beiden Konstruktionen unabhängig voneinander und es besteht dort allenfalls eine gewisse Verwechslungsgefahr – jedoch sicher nicht in Sätzen wie in (80c).

(80) a. daß wir durch den Wald durch gekrochen sind

b. daß wir durch den Wald durchgekrochen sind

c. Durch den Wald durch sind wir gekrochen.

Der Unterschied in der (Nicht-)Inkorporierbarkeit von nachgestellten Präpositionen wird bei Zifonun u. a. (1997, 2086) für eine Unterklassifizierung verwendet. Kombinationen wie in (79a) wird immerhin eine Grammatikalisierung in Richtung Zirkumposition zugestanden.[34] Bei Kombinationen wie in (79b) hingegen wurde in der Literatur (z.B. Engelen (1978)) die Vorstellung vertreten, dass das nachgestellte „durch“ analog zu „hindurch“ als Adverb aufzufassen ist, das von der linksadjazenten PP modifiziert wird. Obwohl die Argumente gegen eine solche Auffassung dargelegt werden in (Zifonun u. a. 1997, 2087) – keine Topikalisierbarkeit wie bei

[33] Tatsächlich finden sich über Google 3 Belege für die Phrase „um des Friedens halber“ (zwei davon in „spontantextlichen“ Forumsbeiträgen). Zudem lassen sich über das Korpuserschliessungswerkzeug unter `http://www.dwds.de` eine Handvoll weiterer Belege aus Literatur- und Zeitungstexten ab 1900 für die Kombination „um...halber“ aufweisen.

[34] Im Grammatikduden (Dudenredaktion 2005, §903) werden „von...ab/an/aus/wegen“ ohne Einschränkung als Zirkumposition beschrieben.

andern Adverbien (81), keine Voranstellbarkeit vor die Präposition (82) – wird keine klare Aussage zum Status dieser Konstruktion gemacht und von „postponierten Präpositionen“ gesprochen.[35]

(81) a. daß wir hindurch durch den Wald krochen.
b. * daß wir durch durch den Wald krochen.

(82) a. Hindurch krochen wir durch den Wald.
b. * Durch krochen wir durch den Wald.

Beispiel (82a) ist meines Erachtens ungeeignet zur Argumentation, da die Lesart mit inkorporiertem Verbpräfix („hindurchkriechen“) zur Verfügung steht. Sobald ein Verb wie in Beispiel (83) verwendet wird, das „hindurch“ nicht inkorporieren kann, ist Topikalisierung schlecht – auch mit „hindurch“.

(83) a. Die Heinzelmännchen entwischten durch den Wald hindurch.
b. * Hindurch entwischten die Heinzelmännchen durch den Wald.

Obige Diskussion sollte die Abgrenzungsproblematik genügend illustriert haben. Zum Abschluss möchte ich die verschiedenen Kriterien, welche in den Argumentationen verwendet worden sind, übersichtlich zusammenstellen und exemplarisch auf den Beispielsatz (85) aus NEGRA anwenden:

1. Wortart des postponierten Elements (Rektionsfähigkeit des postponierten Elements, d.h. kann es überhaupt als Präposition fungieren)
2. Weglassbarkeit des postponierten Elements
3. Weglassbarkeit der PP
4. Verschiebbarkeit des postponierten Elements (vor die PP oder Topikalisierung)
5. Inkorporierbarkeit des postponierten Elements ins Verb
6. Paradigmatische Variabilität der prä- bzw. postponierten Elemente

Zu 1: „herum“ hat klar keine Lesart als Präposition. Allerdings ist dieses Kriterium immer problematisch, weil es gerade distributionell unterschiedliche Phänomene, d.h. Präposition vs. Letztglied von Zirkumposition, durch ein Lexem realisiert haben will. Falls dieses Kriterum für positive Evidenz sorgen soll, ist „am...entlang“ als Zirkumposition zu werten. Denn es verfügt, wie in (84) gezeigt, über eine präpositionale Lesart. Falls es als Ausschlusskriterium dienen soll, wird damit die prototypische Zirkumposition „um..willen“ ausgeschlossen, denn „willen“ alleine kann nicht als Präposition fungieren.

[35] Eine weitere Theorie wird in Olsen (1999) verfochten, welche die Struktur [$_{PP}$ [$_{PP}$ durch den Wald] [$_{P}$ durch]] annimmt. Dabei regiert die nachgestellte Präposition eine linksadjazente PP.

(84) a. entlang der Strasse
b. die/der[36] Strasse entlang
c. an der Strasse entlang
d. entlang an der Strasse

Zu 2: Das postponierte Element „herum“ ist in (85) problemlos weglassbar:

(85) Derweil konnten sich die Eltern im und um das kleine Bierzelt ~~herum~~ bei Speis' und Trank die Zeit vertreiben. [N_{11109}]

Zu 3: Die PP kann nicht weggelassen werden:

(86) * Derweil konnten sich die Eltern im und herum bei Speis' und Trank die Zeit vertreiben.

Zu 4: In der lokalen Lesart ist Voranstellung und Topikalisierung nicht möglich. Allenfalls etwas besser geht es in der direktionalen Leseart[37] in (87c), wobei dort keine Inkorporation ins Verb „verfolgen“ möglich ist. Direktionales „herum“, das inkorporierbar ist und somit auch ohne Präpositionalphrase auftreten kann, ist akzeptabel.

(87) a. * Herum konnten sich die Eltern die Zeit um das Bierzelt vertreiben.
b. * Herum um das Bierzelt konnten sich die Eltern die Zeit vertreiben.
c. ? Herum um das Bierzelt verfolgten sich die Kinder gegenseitig.
d. Herum um das Bierzelt springen die Kinder unermüdlich.

Zu 5: Inkorporierung von „herum“ ist mit dem Verb „vertreiben“ nicht möglich. Da sie bei andern Verben durchaus möglich ist, hat dieses Kriterium mehr damit zu tun, welches Verb abtrennbare Präfixe erlaubt, als mit dem Status von Zirkumpositionen.
Zu 6: Zu „um...herum“ gibt es meines Wissens keine Variation ausser dem Weglassen von „herum“ oder dem Anfügen von „rund“.

Konstituenz von Zirkumpositionen unter Koordination Wegen der flachen Annotation von Präpositionalphrasen kann die Konstituenz des Letzglieds von potentiellen Zirkumpositionen in nicht-koordinierten Kontexten unterspezifiziert bleiben. Wenn aber Koordination hinzukommt, entsteht eine Ungereimtheit, da das öffnende Element sich mit der andern Präposition zu einer Konstituente verbindet und das Letztglied wie in Abbildung 2.1 auf der nächsten Seite aussen vor bleibt.

Wenn man die Konstituenz der Zirkumposition strukturell repräsentieren will, reichen weder die bestehenden Phrasen- bzw. Gruppierungskategorien noch die

[36] Obwohl die normative Grammatik für postponiertes „entlang“ nur Akkusativ-Rektion erlaubt, finden sich viele 'spontanschriftliche' Belege im WWW, welche Dativ-Formen verwenden.

[37] Vgl. dazu auch den Detailtext „Präpositionen als Adverbien“ im Grammis unter `http://hypermedia.ids-mannheim.de`, wo sogenannte Direktionaladverbialia wie in „Er grinste vom Garten herüber“ diskutiert werden.

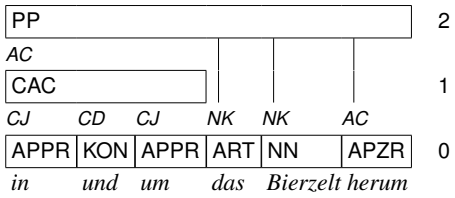


Abbildung 2.1: Originale Annotationsstruktur für „um ... herum“ in NEGRA

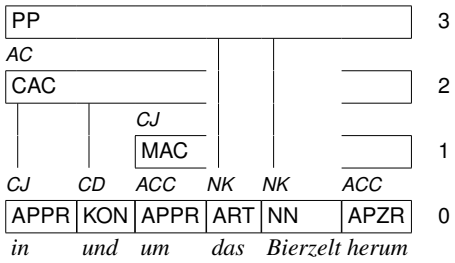


Abbildung 2.2: Vorgeschlagene alternative Annotationsstruktur für „um ... herum“ mit MAC

Funktionsbestimmungen von NEGRA aus. Denn CAC kann auf Grund der Annotationskonventionen für Koordinationen unmittelbar nur die koordinierten Bestandteilen dominieren über CD- und CJ-Funktionen. Wenn man diese Konvention nicht brechen will, braucht man eine Gruppierungskonstituente für Zirkumposition. In Abbildung 2.2 verwende ich dafür in Anlehnung an bestehende NEGRA-Konstrukte MAC (*multi-token adpositional case marker*), sowie das Funktionslabel ACC (*adpositional case marker component*). Die Nachteile der expliziten strukturellen Konstituenz der Zirkumposition werden dabei offensichtlich: Tiefere Verschachtelung sowie eine diskontinuierliche Konstituente.

Erweiterung von CAC mit Partikeln Der Bedeutungskontrast zweier koordinierter Präpositionen kann wie in (88a) noch mit einer Fokuspartikel verstärkt werden. Die erweiterte Präposition „wegen“ wird dort in Ermangelung einer eigenen Phrasenkategorie für solche Spezialfälle als Kopf (HD) einer Adverbialphrase annotiert und die Partikel als Modifikator (MO) davon betrachtet. Damit ist CAC in NEGRA nicht bloss eine Kategorie der Wortkoordination, sondern auch der Wortgruppenkoordination[38]. Die Verwendung der Gedankenstriche in (88a) gibt dem darin koordinierten Teil einschubhaften Charakter, dieser ist aber trotzdem syntaktisch integriert und keine eigentliche Parenthese (vgl. (Dudenredaktion 2005, §1645)).

(88) a. [...] [$_{PP}$ [$_{CAC-AC}$ [$_{APPR}$ Trotz] - [$_{KON}$ oder] [$_{AVP}$ [$_{ADV-MO}$ gerade] [$_{APPR-HD}$ wegen]]] - [$_{PDAT}$ dieser] [$_{PIDAT}$ wenigen] [$_{NN}$ Figuren]]

[38]Im TIGER-Korpus findet sich mit „dies- und jenseits“ in Satz 24’711 zudem eine Morphemkoordination.

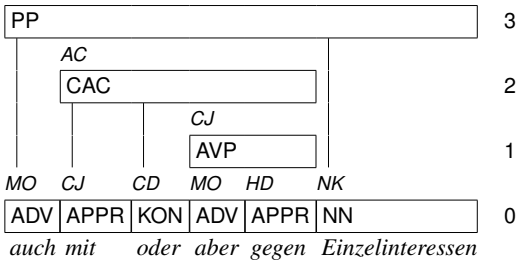


Abbildung 2.3: Originale Annotationsstruktur für „oder aber“ in NEGRA

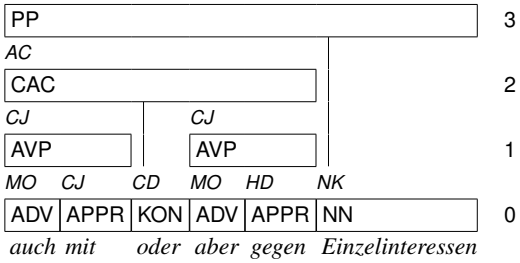


Abbildung 2.4: Vorgeschlagene alternative Annotationsstruktur I für „oder aber“

gehörte der Wiener Walzer zu den schwierigsten Disziplinen: [N_{10112}]

b. [...] die eine laufende Konsensbildung [$_{PP}$ [$_{ADV}$ auch] [$_{PP}$ [$_{CAC-AC}$ mit [$_{KON}$ oder] [$_{AVP}$ [$_{ADV-MO}$ aber] [$_{APPR-HD}$ gegen]] Einzelinteressen]] ermöglicht. [N_{1260}]

In Satz (88b) steht vor beiden Präpositionen ein Modifikator. In Abbildung 2.3 ist die ursprüngliche Annotation der PP im NEGRA-Korpus graphisch dargestellt und man sieht, dass die beiden Modifikatoren strukturell nicht gleich behandelt werden. In Abbildung 2.4 ist eine alternative Annotationsstruktur dargestellt, welche beide Modifikatoren gleich behandelt und somit eine symmetrischere Koordination ergibt.

Man kann sich jedoch fragen, ob die beiden Modifikatoren tatsächlich die gleiche Funktion haben. Der Kontrast in (89c), der sich bei neutralen Betonungsverhältnissen ergibt, spricht dagegen:

(89) a. [...] wird es vor allem darauf ankommen, eine vergleichbare organisatorische Infrastruktur aufzubauen, die eine laufende Konsensbildung auch mit oder aber gegen Einzelinteressen ermöglicht. [N_{1260}]

b. [...] wird es vor allem darauf ankommen, eine vergleichbare organisatorische Infrastruktur aufzubauen, die eine laufende Konsensbildung auch mit ~~oder aber gegen~~ Einzelinteressen ermöglicht.

c. * [...] wird es vor allem darauf ankommen, eine vergleichbare organisatorische Infrastruktur aufzubauen, die eine laufende Konsensbildung ~~auch mit oder~~ aber gegen Einzelinteressen ermöglicht.

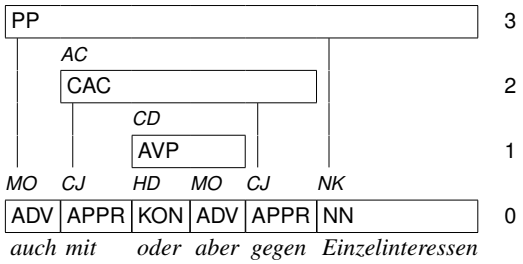


Abbildung 2.5: Vorgeschlagene alternative Annotationsstruktur II für „oder aber“

Das Wort „aber“ scheint stärker mit „oder“ verknüpft als mit „gegen“ – dieser Befund wir auch vom Grammatik-Duden (Dudenredaktion 2005, §1762) unterstützt, der die Wortkombination „oder aber“ unter dem Titel „Alternative Konnektoren mit exklusiver Lesart“ behandelt. Die Annotation von Koordinationen, wo der Konjunktor aus einer Wortgruppe besteht, wird in NEGRA unterschiedlich annotiert:

(90) a. [...] war zwar keine Kirche, [$_{AVP}$ [$_{ADV-MO}$ wohl] [$_{KON-HD}$ aber]] eine mittelalterliche Wehranlage [...] [N_{2073}]

b. [...] den tadschikischen Partisanen Massuds, [$_{AVP}$ [$_{KON-AVC}$ geschweige] [$_{ADV-AVC}$ denn]] den usbekischen Milizionären. [N_{11433}]

In (90a) wird eine Kopf-Modifikator-Abhängigkeit in der AVP gesehen – motiviert durch die Weglassbarkeit des Modifikators. In (90b) dagegen wird „geschweige denn“ als Mehrwortlexem aufgefasst, dessen Teile nicht in eine Abhängigkeit gebracht werden. Dies, obwohl „geschweige“ gut als selbstständiger Koordinator funktionieren kann, und somit im Einklang mit dem KON-Tag als Kopf zu analysieren wäre.

Weitere Annotationsmöglichkeiten, welche sich in der Rolle von „auch“ unterscheiden, möchte ich an dieser Stelle nicht diskutieren. Denn die Extension und der exakte syntaktische Status von sogenannten Fokus-Partikeln vor NP/DP und PP ist in der theoretischen Linguistik stark umstritten. Insbesondere die Wiederaufnahme und Weiterentwicklung der These von Jacobs (1983) durch Büring und Hartmann (2001), dass solche Partikel immer verbangebunden sein müssen und nicht adnominal analysiert werden dürfen, hat so fundamental unterschiedliche Dominanzverhältnisse in der Satzstruktur zur Folge, dass hier keine theorieneutrale Unterspezifikation mehr möglich ist. Eine ausführliche kritische und falsifikationsorientierte Diskussion dieser These findet sich in Reis (2005). Eine komparatistische und quantitative Erörterung bieten Bouma u. a. (2007).

CAC vs. CAVP Nebst den CAC-Knoten, welche präpositional funktionieren und somit immer das Funktionsetikett AC tragen, gibt es im NEGRA-Korpus noch einen Vertreter mit Modifikatorfunktion (MO) wie in Beispiel (91a).

in %	Anzahl	Funktion	Tochterkonstituenten
57.1	8	OC	VVPP KON VVPP
7.1	1	PD	VVPP VVPP VVPP
7.1	1	OC	VVPP VVPP VVPP
7.1	1	OC	VVPP VVPP
7.1	1	OC	VVINF KON VVINF
7.1	1	NK	VVINF KON VVINF
7.1	1	HD	VVPP KON VVPP

Tabelle 2.38: Die 14 Fälle von CVP als Wortkoordination in NEGRA

(91) a. [$_{CAC-MO}$ [$_{APPR}$ Nach] [$_{KON}$ und] [$_{APPR}$ nach]] stellte sich heraus [...] [N_{8982}]

b. [...] daß man dann [$_{CAVP-MO}$ [$_{ADV}$ nach] [$_{KON}$ und] [$_{ADV}$ nach]] Erfahrungen mit anderen, problematischeren Stoffen sammelt". [N_{19570}]

Die restlichen 7 Vorkommen von „nach und nach“ sind hingegen alle entsprechend dem Muster in (91b) als CAVP annotiert. Bei (91a) liegt eine falsche Annotation vor, welche in der nicht aufgelösten Homonymie der Präposition „nach“ mit dem Adverb „nach“ auf der Ebene der Wortartenkategorie ihren Anfang nimmt und dann erst auf der syntaktisch-funktionalen Ebene desambiguiert wird.

Dieser Fall ist insofern speziell, weil „nach und nach“ ein reduplizierend koordiniertes Mehrwortlexem dargestellen, das analog in „durch und durch“, „über und über“ (im Sinn von „vollständig“) oder „für und für“ (veraltet für „immer“) zu finden ist.

Was CAC und was CAVP sein soll, lässt sich eigentlich problemlos entscheiden, wenn die präpositionale Rektion für CAC als Kriterium genommen wird.

CAC als elliptische CPP Der Kategorie CAC, welche prototypisch als Wortkoordination realisiert wird, könnte durchaus als kataleptische CPP aufgefasst werden, wo bei der ersten Präposition die eingebettete Nominalphrase weggelassen wurde. Diese Annotationsform ist meines Wissens für Deutsch nicht gemacht worden, auch wenn es im Grammatik-Duden (Dudenredaktion 2005, §1424) heisst: „Bei einer Reihung mit unterschiedlichen Präpositionen müssen gleiche Nominalphrasen nur einmal gesetzt werden“.

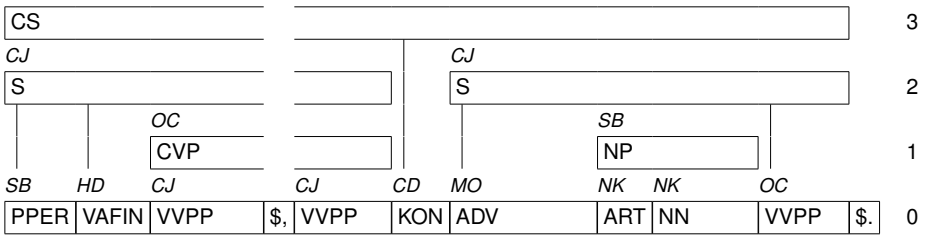
2.4.6 CVP

Wie in Tabelle 2.43 auf Seite 101 zu sehen, sind auf der Wortebene koordinierte CVP erstaunlich selten in NEGRA, es gibt nur 14 Fälle. Die Tabelle 2.38 zeigt, dass es sich dabei fast ausschliesslich um koordiniertes Partizip Perfekt handelt. Die Frage ist, ob dies der sprachlichen Realität entspricht oder aus Annotationskonvention geschieht.

Grundsätzlich würde man von einer interpretierten syntaktischen Annotation wie in NEGRA, welche die Abhängigkeiten der Satzglieder ausdrücken will, erwarten, dass sie dies im Fall der koordinierten Elemente so tut, dass alle syntaktischen Funktionen, welche von einem Konjunkt wahrgenommen werden, strukturell transparent kodiert sind. Im Fall der CS wurde dies zwar wie schon erwähnt explizit nicht so gemacht, für die nicht-finiten CVP fehlen jedoch solche Angaben.

Das Problem der zu engen Koordinationsannotation Im Beispiel (92a) aus NEGRA ist eine asyndetische CVP mit „kontrolliert und angezeigt“ annotiert, welche vom Hilfsverb „wird“ abhängig ist. Der zweite Teilsatz ist dagegen ohne eine finites Verb aufgebaut – die Abhängigkeit des Partizips „verhängt“ vom Hilfsverb „werden“ ist nicht strukturell transparent kodiert (und zudem ist ein falsches Subjekt gesetzt). D.h. die Koordination wurde zu eng annotiert.

(92) a.



[N_{3039}]

b. Es wird [$_{CVP-OC}$ [$_{VVPP-CJ}$ kontrolliert] , [$_{VVPP-CJ}$ angezeigt] und [$_{VP-CJ}$ gegebenenfalls [$_{NP-OA}$ eine Geldstrafe] [$_{VVPP-HD}$ verhängt]]].

In (92b) ist dargestellt, wie die Abhängigkeiten durch das Annotieren einer monosyndetischen CVP transparent kodiert werden.

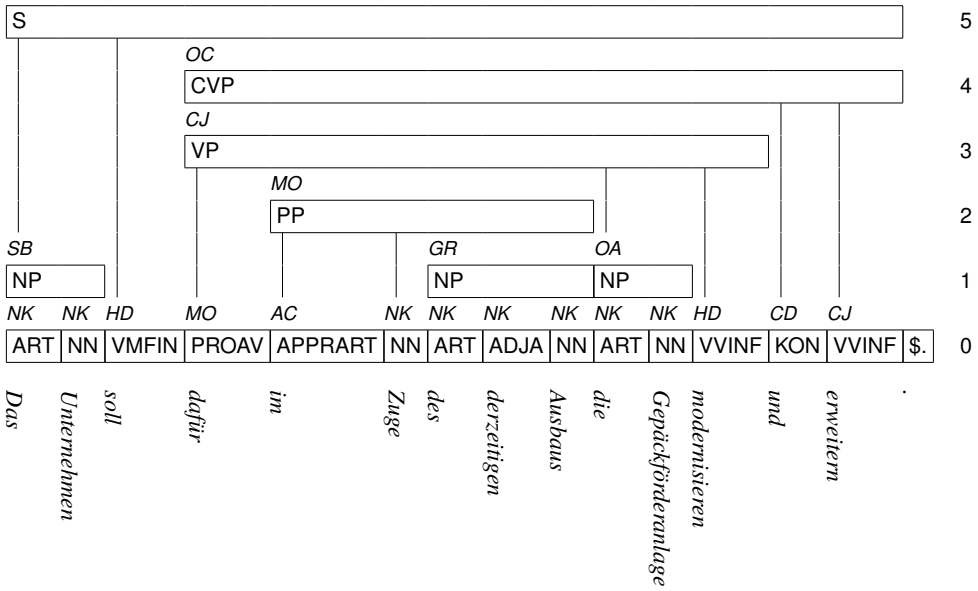
Das Problem der gierigen Konjunkte Zu enge Annotation ist nur ein kleines Problem in NEGRA. Viel öfter wird wie im Beispiel (93a) strukturiert, wo das Erstglied als vollständige CVP annotiert ist. Alle abhängigen Objekte und Adverbiale sind diesem Erstglied zugeordnet. Das zweite Konjunkt steht dann strukturell isoliert von den Objekten und Adverbialen, welche es semantisch aber ganz klar integriert. Dies entspricht keineswegs den Intentionen eines semantisch interpretierten Korpus.

Der Sinn von Satz „Das Unternehmen soll dafür im Zuge des derzeitigen Ausbaus die Gepäckförderanlage modernisieren und erweitern“ ist nicht die Paraphrase „Das Unternehmen soll dafür im Zuge des derzeitigen Ausbaus die Gepäckförderanlage modernisieren. Das Unternehmen soll erweitern“, welche durch die Annotation in (93a) nahegelegt wird.

Eine Annotation, welche die Paraphrase (93b) strukturell transparent umsetzt, ist in (93c) skizziert. Sie verlangt, dass die Infinitive eng auf der Wortebene koordiniert werden. Beide Modifikatoren und das Akkusativobjekt sind dabei abhängig

vom koordinierten Verbalkopf (HD). Dabei ist die eigentlich einfache und natürliche Forderung erfüllt, dass die koordinierten Elemente gegen Aussen dieselben thematischen Abhängigkeiten und damit einhergehend dieselbe Kasusforderung aufweisen.

(93) a.



[N_{842}]

b. Das Unternehmen soll dafür im Zuge des derzeitigen Ausbaus die Gepäckförderanlage modernisieren. Das Unternehmen soll dafür im Zuge des derzeitigen Ausbaus die Gepäckförderanlage erweitern.

c. Das Unternehmen soll [$_{VP-OC}$ [$_{PROAV-MO}$ dafür] [$_{PP-MO}$ im Zuge des derzeitigen Ausbaus] [$_{NP-OA}$ die Gepäckförderanlage] [$_{CVP-HD}$ [$_{VP-CJ}$ modernisieren] und [$_{VP-CJ}$ erweitern]]] .

Die Beispiele in (94) zeigen, dass unabhängig von der Reihenfolge der Konjunkte deren Kasusforderung erfüllt sein muss. In (Sternefeld 2005, H1) wird dies als *external uniformity condition* angesprochen.

(94) a. * Fritz hat mir/mich geholfen und hochgehoben.

b. * Fritz hat mir/mich hochgehoben und geholfen.

c. Fritz hat mich gestützt und hochgehoben.

d. Fritz hat mir geholfen und vertraut.

Solche koordinierten komplexe Prädikate teilen ihre gemeinsamen Argumente syntaktisch und semantisch. Sternefeld erwähnt ein Beispiel aus dem *Hohlspiegel* vom 25.8.2003, wo dieser Effekt auf Grund einer milden Form von Holzweganalyse („garden path sentence“) unbeabsichtigtes *structure sharing* auslöst:

(95) a. Die SPD grillt und ehrt ihre jahrelangen Mitglieder.
b. , dass die SPD grillt und ihre jahrelangen Mitglieder ehrt.
c. Die SPD grillt am Abend und ehrt ihre jahrelangen Mitglieder.
d. Am Abend grillt die SPD und ehrt ihre jahrelangen Mitglieder.

Beim Lesen des Satzes (95a) fasst man „grillt und ehrt“ als verbale Wortkoordination auf. In Übereinstimmung mit der *external uniformity condition* teilen sie sich syntaktisch (und semantisch) das Subjekt und das Objekt.

Sobald „grillt“ und „ehrt“ nicht mehr Kontaktstellung aufweisen, wird die intendierte Lesart prominent, welche auf das Grillieren von Mitgliedern verzichtet. Ob dies durch die Verbletztstellung (95b), Temporaladverbien (95c) oder in Kombination mit einer Subjektlücken-Konstruktion[39] (95d) geschieht, ist einerlei: Das Objekt ist in allen Fällen nicht mehr von beiden Prädikaten gleichzeitig benutzt.

Ob durch Konjunktoren vermittelte Kontaktstellung von Verben vorliegt oder nicht, weist somit einen starken Einfluss auf die syntaktische Struktur und die Interpretation auf, was die Vermutung erlaubt, dass solche Konstruktionen wie in (95) strukturell nicht gleich behandelt werden sollten.

Die Betonung der Gleichförmigkeit nach Aussen in der *external uniformity condition* ist in Bezug auf das zu sehen, was im Innern von Koordinationen gleich bzw. nicht gleich sein muss. So betont Sternefeld, dass im Innern nicht die Gleichförmigkeit auf der Ebene der lexikalischen bzw. daraus projizierten syntaktischen Kategorie zählt, sondern dass die „semantische Funktion“ entscheidend ist. Sternefeld (2005, H1) gibt dazu die Beispiele:

(96) a. Sie sang [$_{AP}$ sehr laut] und [$_{PP}$ ohne Vibrato].
b. Er war [$_{DP}$ ein Idiot] und [$_{AP}$ stolz darauf].

Was mit dem Kriterium „semantische Funktion“ genau gemeint ist, führt Sternefeld nicht weiter aus. Es kann aber zumindest in den obigen Fällen gut durch die klassischen strukturalistischen Fragetests ersetzt werden, welche die Komplemente sowie die Adverbiale in ihrer semantischen Funktion identifizieren.

(97) a. Wie sang sie? [$_{AP}$ Sehr laut] und [$_{PP}$ ohne Vibrato].
b. Was war er? [$_{DP}$ Ein Idiot] und [$_{AP}$ stolz darauf].

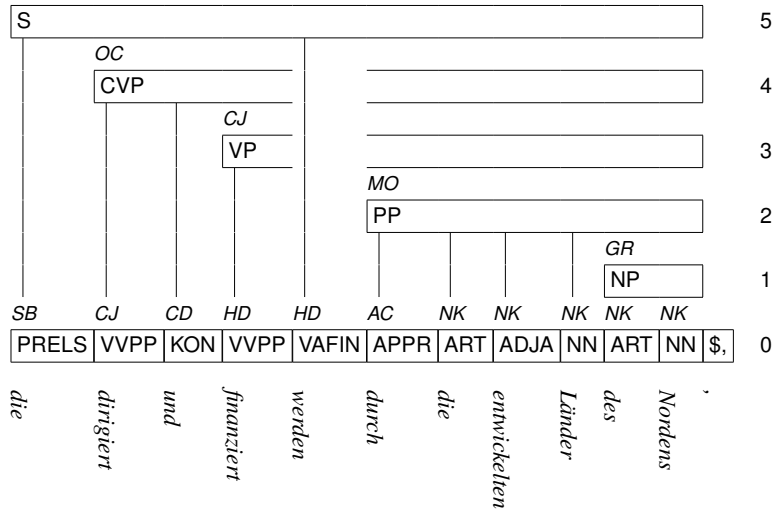
Gierige Letztkonjunkte Gierige Konjunkte treten aber nicht bloss als Erstglieder auf in der Annotation. Es besteht eine Tendenz, alle abhängigen Glieder nur an dasjenige Konjunkt zu hängen, welches nicht durch ein Komma oder einen Konjunktor abgetrennt ist. Der eingebettete Relativsatz in (98) bindet das Passivsubjekt ganz klar sowohl an „dirigiert“ wie an „finanziert“, aber die Annotation modelliert diese Abhängigkeit für „dirigiert“ nicht. Die Wortkoordination bleibt somit versteckt.

[39] Siehe dazu Höhle (1983), Fortmann (2005) sowie Sternefeld (2005, Abschnitt Asymmetrische Koordination).

Tag	Anzahl	in %	Typ	Anzahl	in %
VVPP	33	41.2	1	18	22.5
			0	13	16.2
			X	1	1.2
			3	1	1.2
VVFIN	26	32.5	1	15	18.8
			2	4	5.0
			3	3	3.8
			X	2	2.5
			0	2	2.5
VVINF	19	23.8	X	8	10.0
			0	7	8.8
			1	4	5.0
VVIMP	2	2.5	X	1	1.2
			0	1	1.2

Tabelle 2.39: Klassifikation der potentiellen Fehlannotationen für CVP in NEGRA

(98)



[N_{15484}]

Evaluation der potentiellen Wortkoordinationen in NEGRA Um genauer abschätzen zu können, wie viele Wortkoordinationen durch problematische Annotation in NEGRA verloren gegangen sind, wurden alle Verbformen mit identischem STTS-Tag gesucht, welche nur durch ein Komma oder einen Konjunktor voneinander abgetrennt sind. Über das ganze NEGRA-Korpus ergab dies insgesamt 155 Kandidaten. Die Tabelle 2.39 schlüsselt für die Fälle aus der ersten Hälfte des NEGRA-Korpus auf, wie die Kandidaten zu bewerten sind.

In der Spalte „Typ“ werden folgende Kodes verwendet: „0“ = Annotierung ist korrekt, „1“ = gieriges Erstglied, „2“ = gieriges Letzglied, „3“ = gieriges Erst- und

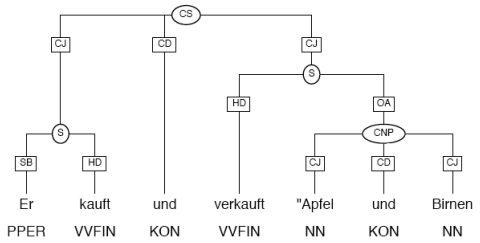


Abbildung 2.6: Die verlangte Annotation für koordinierte finite Verbformen in NEGRA (Brants u. a. 1999, 88)

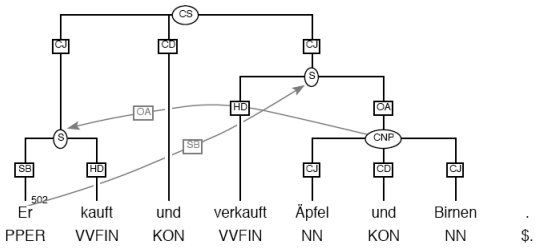


Abbildung 2.7: Die verlangte Annotation für koordinierte finite Verbformen in TIGER (Albert u. a. 2003, 118)

Letzglied, „X“ = nicht anwendbar. Der Typ „X“ wurde bei der Auswertung immer gewählt, wenn die syntaktische Konstruktion nicht eindeutig für eine Integration gesprochen hat. Dies war insbesondere bei Reihungen der Fall mit einer syntaktisch starken Absetzung. Nur auf die Typen bezogen ergibt sich folgende Auswertung: Typ 1 (37, 46.2%), 0 (23, 28.8%), X (12, 15.0%), 3 (4, 5.0%), 2 (4, 5.0%). Problematische Annotationen sind also deutlich häufiger zu finden als korrekte.

Koordinierte finite Verben in NEGRA und TIGER Eine besondere Schwierigkeit mit NEGRA ergibt sich bei koordinierten finiten Verben. Während es für die nicht-finiten Verbalformen mit der VP und der CVP ausreichende Beschreibungsmöglichkeiten für koordinierte Konstruktionen gibt, fehlt bei den finiten Verbalformen eine koordinierende Kategorie – ausser man wollte tatsächlich CS dazu verwenden. Im Handbuch von NEGRA (Brants u. a. 1999, 88) wird deshalb die in Abbildung 2.6 gezeigte Annotation bestimmt. Dabei fehlt genau wie oben ausführlich diskutiert eine strukturell transparente Kodierung aller syntaktischen Abhängigkeiten.

Das Caveat „Die Annotation von Koordinationen ist vorläufig!“ im NEGRA-Handbuch (Brants u. a. 1999, 86) wird vermutlich auch damit zusammenhängen.

Bei der Annotation von Koordinationen tritt beim TIGER-Korpus, welches durchaus als Nachfolgekorpus verstanden werden kann, eine Änderung ein. Die

fehlende transparente Kodierung aller syntaktischen Abhängigkeiten bei Koordinationen wird dort mit Hilfe von sekundären Kanten gelöst. Die Abbildung 2.7 zeigt den gleichen Satz gemäss Annotation in TIGER (Albert u. a. 2003, 118). Die syntaktischen Abhängigkeiten sind damit explizit gemacht, allerdings nicht auf der strukturell primären Ebene. Dies hat den Vorteil, dass TIGER und NEGRA auf der primären Strukturebene kompatibel bleiben. Die zusätzliche Annotationsebene erlaubt es auch, Konstruktionen wie „Gapping“ mit den gleichen Mitteln zu erfassen. Es stellt sich trotzdem die Frage, ob eine solch lokale Koordination wie in Abbildung 2.7 auf der vorherigen Seite auf der primären Strukturebene nicht adäquater zu beschreiben wäre. Das Einzige, was es dazu braucht, ist eine Koordinationskategorie für finite Verbalformen wie in Beispiel (99).

(99) [$_S$ er [$_{CVVFIN-HD}$ kauft und verkauft] [$_{CNP}$Äpfel und Birnen]].

2.4.7 CVZ

Die Koordinationskonstituente CVZ (*coordinated zu-marked infinitive*) dominiert nur VZ (*zu-marked infinitive*). Allerdings umfasst VZ nicht bloss die Partikel „zu“ in Kombination mit einem Infinitiv, dem sogenannten „zu“-Infinitiv. Das seltenere attributive Partizip I mit „zu“ wird ebenfalls mit VZ ausgezeichnet.

Wegen ihrer unterschiedlichen Distribution diskutiere ich die beiden Verwendungsweisen in eigenen Abschnitten. Im Abschnitt 2.4.7.2 wird zudem gefragt, warum überhaupt VZ bzw. CVZ als syntaktischer Knoten annotiert ist und ob dies eigentlich eine konsequente Umsetzung der NEGRA-Annotationsstrategie darstellt. Gemäss Tabelle 2.40 auf der nächsten Seite treten CVZ selten auf. Zudem wird knapp die Hälfte der VZ in koordinierender Funktion in CVP integriert. Dies geschieht dann, wenn das Verb noch erweitert ist um Objekte.

2.4.7.1 „zu“-Infinitiv

Obwohl die Konjunkte nur bei Verben mit abtrennbarem Verbpräfix VVIZU durch ein einzelnes Token repräsentiert sind, soll CVZ hier vollständig unter Wortkoordination diskutiert werden. Ich halte mich damit an den Grammatik-Duden (Dudenredaktion 2005, § 611), für den das Wort „zu“ vergleichbar zum „ge“ des Partizips II als ein vorangestelltes Flexionselement gilt. In der GDS-Grammatik (Zifonun u. a. 1997, 2159) wird „zu“ mit derselben Begründung auch bei orthographischer Abtrennung als Verbaffix bezeichnet.

Zwischen „zu“ und den Infinitiv kann kein lexikalisches Material treten und bei Koordination zweier Infinitivformen muss „zu“ gemäss Duden wiederholt werden. Beispiel (100b) mit einem Akkusativobjekt ist ganz klar ungrammatisch. Ohne das Objekt wie in (100c) scheint mir der Verstoss deutlich milder. Die Akzeptabilität steigt weiter, wenn wie im Beispiel (100d) aus dem TIGER-Korpus eine idiomatische Verbkoordination verwendet wird.

Kategorie	Anz.	in %	Funk.	Anz.	in %	Mutter	Anz.	in %
VZ	1695	75.8	HD	1620	72.5	VP	1602	71.7
						AP	18	0.8
			OC	42	1.9	S	33	1.5
						VP	4	0.2
						NP	4	0.2
						AP	1	0.0
			CJ	17	0.8	CVZ	9	0.4
						CVP	8	0.4
			NK	9	0.4	NP	5	0.2
						PP	4	0.2
			MNR	6	0.3	NP	6	0.3
			RE	1	0.0	NP	1	0.0
VVIZU	540	24.2	HD	519	23.2	VP	519	23.2
			OC	12	0.5	S	4	0.2
						VP	3	0.1
						NP	3	0.1
						AP	2	0.1
			CJ	7	0.3	CVP	6	0.3
						CVZ	1	0.0
			RE	1	0.0	NP	1	0.0
			PD	1	0.0	S	1	0.0

Tabelle 2.40: Verhältnis der Kategorien VZ und VVIZU zu ihrer Mutterkategorie in NEGRA. Legende: Funk. = Funktion bezüglich der Mutterkonstituente. Lesebeispiel: Von 17 Vorkommen von VZ fungieren 9 als Konjunkt (CJ) innerhalb einer CVZ und 8 innerhalb einer CVP.

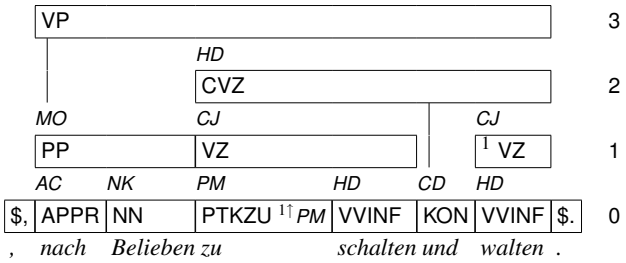


Abbildung 2.8: Ausschnitt aus der TIGER-Annotation von Satz 16213

in %	Anzahl	Tochterkonstituenten
60.0	15	VZ KON VZ
24.0	6	VZ KON VVIZU
12.0	3	VVIZU KON VZ
4.0	1	KON VZ VZ KON VZ

Tabelle 2.41: Verteilung der CVZ in NEGRA und TIGER

(100) a. Ich muss Ihnen davon abraten, täglich zu rauchen und Alkohol zu trinken.

b. * Ich muss Ihnen davon abraten, täglich zu rauchen und Alkohol trinken.

c. ? Ich muss Ihnen davon abraten, täglich zu rauchen und trinken.

d. ? Seine Partei wolle dem Präsidenten nur nicht erlauben, nach Belieben zu schalten und walten. [T_{16213}]

Im TIGER-Korpus ist beim Satz 16213 das fehlende „zu“ beim zweiten Konjunkt durch eine sekundäre Kante als elliptisch annotiert (siehe Abbildung 2.8). Wenn man in solchen Fällen keine analeptische Konstruktion sehen will und eine syntaktische Struktur zuweisen möchte, bleibt noch die Möglichkeit, ein „zu“ mit koordiniertem Infinitiv anzunehmen. Wegen dem Skopus von „zu“ über eine syntaktische Koordination ergeben sich jedoch Schwierigkeiten, dieses „zu“ als morphologisches Affix zu betrachten. Allerdings zeigt das Beispiel gerade auch, dass die Akzeptabilität an idiomatische Wendungen geknüpft erscheint, welche unbestritten eine lexikalisierte Erscheinung darstellen.

Den 1686 nicht-koordinierten Vorkommen von VZ (*zu-marked infinitive*) stehen im NEGRA-Korpus gerade mal 7 koordinierte Vorkommen von VZ gegenüber. Den 539 nicht-koordinierten Vorkommen von VVIZU (Vollverb im Infinitiv mit „zu“) steht 1 einziges koordiniertes Vorkommen gegenüber. Wegen der dürftigen Datenlage wurden auch die Vorkommen im TIGER-Korpus erhoben; dort finden sich 8 koordinierte vs. 1243 nicht koordinierte VVIZU sowie 33 koordinierte vs. 3590 nicht koordinierte VZ.

2.4.7.2 „zu“-Partizip I

Nebst der Form „zu“ + Infinitiv gibt es noch die Kombination „zu“ + attributives Partizip I wie in (101a). Wie Beispiel (101b) zeigt, besteht ein diathesenähnliches Verhältnis zwischen der pränominalen attributiven Gerundiv-Konstruktion und einem Relativsatz, welcher „sein zu + Infinitiv“ als „passivisches Modalitätsverb“ (Dudenredaktion 2005, § 579; § 829) enthält.

(101) a. Soviel er dem Kunstwerk an Ewigkeitswerten entzieht, soviel räumt er ihm ein an Bedeutung für den human zu lebenden (und zu überlebenden) Tag. [N_{2432}]

b. [...] räumt er ihm ein an Bedeutung für den Tag, der human zu leben und zu überleben ist.

Um diesen verbalen Bezug transparent zu kodieren, werden im NEGRA-Korpus die knapp 30 attributiven „zu“-Partizipien I konsequent annotiert wie in (102a).

(102) a. [$_{VZ-NK}$ [$_{PTKZU-PM}$ zu] [$_{ADJA-HD}$ *Adjektiv*]]

b. [$_{AP-NK}$ [...] [$_{PTKZU-PM}$ zu] [$_{ADJA-HD}$ *Adjektiv*]]

Im TIGER-Korpus hingegen findet sich in knapp der Hälfte die andere ebenfalls naheliegende Annotation als Adjektivphrase wie in (102b).

Diese Inkonsistenz mag damit zu tun haben, dass die Annotationskonvention im TIGER-Annotationsschema in der älteren Version (Brants u. a. 2000, 33) AP vorschreibt, in der neueren Version (Albert u. a. 2003, 43) hingegen VZ. Es widerspiegelt aber auch die Tatsache, dass es sich bei deverbalen Adjektiven um lexikalisch komplexe Kategorien handelt.

Wenn die Konstituenten-Kürzel VZ bzw. CVZ nicht nur für Infinitiv-Konstruktionen, sondern auch für attributive Partizipien verwendet werden, entsteht mit dieser Annotationsstrategie unter Koordination eine unschöne lokale Mehrdeutigkeit über zwei syntaktische Ebenen. Wie aus Abbildung 2.9 auf der nächsten Seite ersichtlich desambiguiert erst die Mutterkategorie von CVZ bzw. der Kopf der Tochter von VZ zwischen „zu“-Infinitiv bzw. -Partizip. Global besteht selbstverständlich keine Mehrdeutigkeit, da das Partizip I attributiv immer als Adjektiv getaggt ist.

Warum VZ? Warum CVZ? Ein wichtiges Annotationsprinzip bei NEGRA und TIGER ist die flache Strukturierung. Die syntaktischen Zusammenhänge werden deshalb weniger durch Verschachtelung von Strukturen ausgedrückt, sondern durch explizites funktionales Kategorisieren der syntaktischen Verhältnisse zwischen den Knoten und ihrem gemeinsamen Mutterknoten. Grundsätzlich spricht nichts dagegen, das morphologische Element gemäss dem Prinzip der flachen Annotation als PTKZU-PM und das dazugehörige Verb ohne Zwischenkategorie VZ einzeln an die Mutterkategorie anzuhängen.

Die Schwierigkeit entsteht, wenn solche Paare koordiniert werden. Ein weiteres wichtiges Annotationsprinzip besagt, dass Konjunkte immer als syntaktische

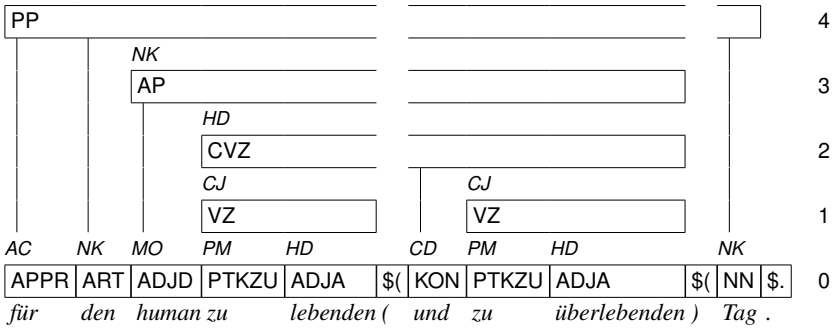


Abbildung 2.9: Ausschnitt aus NEGRA-Annotation von Satz 2432

Einheit mit den Konjunktoren verknüpft werden. Im Fall von CVZ muss für die Kombination „zu“ + Infinitiv somit sowieso eine Zwischenkategorie eingefügt werden. Wenn man dazu nicht die Phrasenkategorie VP wiederverwenden will, braucht es tatsächlich eine eigene Konstituente.

2.4.8 CCP

Diese Kategorie kommt sowohl im NEGRA- wie im TIGER-Korpus offiziell nur ein einziges Mal vor.

(103) a. [$_{CCP-CP}$ [$_{KOUS}$ Obwohl] [$_{KON}$ oder] [$_{AVP}$ [$_{ADV-MO}$ gerade] [$_{KOUS-HD}$ weil]]] soviel Bewegung und Umbau auf der Bühne ist. [N_{17092}]

b. [$_{CCP-CP}$ [$_{KOUS}$ Da] [$_{KON}$ und] [$_{KOUS}$ wenn]] die Befriedigung unseres Modernisierungs- und Strukturwandlungsbedarfs [...] [T_{17674}]

Wegen der Fokuspartikel „gerade“ liegt bei (103a) aus dem NEGRA-Korpus allerdings keine als solche annotierte Wortkoordination vor.[40] Die Diskussion von CCP soll aber trotzdem an dieser Stelle stattfinden, da unterordnende Konjunktionen (KOUS) sowohl in traditionelleren wissenschaftlichen wie in modernen theoretisch-generativen Grammatiken eine wichtige wortbezogene Rolle spielen.

Der Status von unterordnenden Konjunktionen in der Duden-Grammatik Im Grammatik-Duden (Dudenredaktion 2005, §1339, §1349ff.) wird eine gängige Stellungsfeldertheorie[41] für deutsche Sätze referiert. Dieses topologisch Modell ist in der Tabelle 2.42 illustriert. Die Füllung der linken Satzklammer wird folgendermassen einschränkt (Dudenredaktion 2005, §1369): „Die linke Satzklammer ist entweder mit genau einer Wortform besetzt – oder mit gar keiner.“

Die Forderung, dass in der linken Satzklammer keine Wortgruppen auftauchen dürfen, verbietet die Konstruktion in (103a) gleich doppelt:

[40] Die Problematik solcher Fokuspartikeln wurde unter 2.4.5 auf Seite 79 schon im Zusammenhang mit CAC angesprochen.

[41] Einen Überblick über verschiedene Varianten von Stellungsfeldertheorien bietet (Haftka 1993).

VF	LSK	MF	RSK
–	Obwohl	soviel Bewegung auf der Bühne	ist
–	Obwohl	soviel Bewegung auf der Bühne	gewesen ist
Soviel Bewegung	ist	auf der Bühne	–
Soviel Bewegung	ist	auf der Bühne	gewesen

Tabelle 2.42: Standardannahmen des topologischen Modells für Deutsch gemäss der Duden-Grammatik: VF=Vorfeld, LSK=Linke Satzklammer, MF=Mittelfeld, RSK=Rechte Satzklammer

- Partikeln dürfen nicht mit der Subjunktion zusammen im gleichen Feld stehen.

- Koordinierte Subjunktionen dürfen als Wortgruppe nicht die linke Satzklammer besetzen.

Die Duden-Grammatik (Dudenredaktion 2005, §1385) erörtert die Probleme nur bezüglich beiordnenden Konjunktionen und Abtönungspartikeln, die sich auf den ganzen Satz oder das ganze Satzgefüge beziehen. Sie schlägt für die grammatische Beziehung dieser Redeteile zum Vorfeld bzw. zur linken Satzklammer den Terminus „Anlehnung“ vor. So lehnt sich das Wort „Und“ in (104a) ans Vorfeld an, das Wort „Doch“ lehnt sich an die linke Satzklammer in (104b). Falls das Vorfeld auch bei Verbletztsätzen als strukturell vorhanden angenommen wird, erscheint die beiordnende Konjunktion somit nach diesem leer angesetzten Feld. Eine Anlehnung ist auch rückwärts gerichtet möglich wie in (104c), wo das Wort „aber“ sich ans linksseitige Vorfeld lehnt.

(104) a. Und [$_{VF}$ Fritzchen] [$_{LSK}$ will] [$_{MF}$ schon wieder mit dem Aschenbecher] [$_{RSK}$ spielen]

b. [$_{VF}$] Doch [$_{LSK}$ als] [$_{MF}$ Fritzchen schon wieder den Aschenbecher] [$_{RSK}$ ausleerte], [...]

c. [$_{VF}$ Der Grossvater] aber [$_{LSK}$ raucht] [$_{MF}$ trotzdem seine Zigarren] [$_{RSK}$].

Zu Fokuspartikeln, wie sie etwa im Beispielsatz (103a) erscheinen, sagt die Duden-Grammatik in diesem Zusammenhang nichts. Nur bei der Diskussion der Wortarten wird unter §873 erwähnt, dass „Fokuspartikeln im Verbund mit ihrer Konstituente den Informationskern (Fokus) des Satzes“ bilden. Der syntaktische Status dieses Verbunds bleibt unbestimmt und die Vorschläge zur syntaktischen Strukturierung in Reis und Rosengren (1997) unbeachtet.

Eine von der Duden-Grammatik abweichende Feldertheorie wird in der Linguistik -Einführung von Meibauer u. a. (2002) propagiert. Dort werden folgende Satzteile ins gleiche Feld genommen: Unterordnende Konjunktionen wie in (105a);

Phrasen mit Relativpronomen wie in (105b) oder indirekt verwendeten Fragewörtern wie in (105c) oder (105d). Eine Kritik dieses Ansatzes findet sich in Sternefeld (2005).

(105) a. Er weiss, [$_{LSK}$ dass] [$_{MF}$ man heute leicht Geld] [$_{RSK}$ verdienen kann]
b. Hier sind die Bücher, [$_{LSK}$ mit denen] [$_{MF}$ man heute leicht Geld] [$_{RSK}$ verdienen kann]
c. Ich weiss, [$_{LSK}$ wie] [$_{MF}$ man heute leicht Geld] [$_{RSK}$ verdienen kann]
d. Ich weiss, [$_{LSK}$ mit welchen Büchern] [$_{MF}$ man heute leicht Geld] [$_{RSK}$ verdienen kann]

Der Status von unterordnenden Konjunktionen in der Generativen Grammatik Die Beschränkung der unterordnenden Konjunktionen auf eine Satzstelle, welche keine Wortgruppen oder Phrasen zulässt, sondern nur Einwortfüllung, findet sich ebenfalls verbreitet in den verschiedenen Grammatikmodellen der Generativen Syntax. Eine gute Übersicht zur Entwicklung und den Zusammenhängen sowie eine kritische Diskussion gibt Lee (1999). Insbesondere seit dem Theoriestand von Chomsky (1986), wo das S/S'-System zur Analyse von Sätzen durch das CP/IP-System im Sinne einer bezüglich Kopf- und Spezifikatorposition einheitlichen X'-Theorie abgelöst wird, ergibt sich für die unterordnenden Konjunktionen eine natürliche Kategorisierung als lexikalischer Kopf C der Komplementiererphrase CP.

Diese Idee haben die Vertreter der sog. Uniformitätsthese, die für Haupt- wie für Nebensätze des Deutschen die gleiche Grundstruktur wie in (106a) bzw. (106b) ansetzen, am konsequentesten durchgesetzt. Dabei wird C als Positionskategorie betrachtet, welche in Nebensätzen durch den Komplementierer gefüllt ist und in Hauptsätzen das finite Verb enthält. Wie durch (106c) deutlich werden sollte, sind sich die Vertreter der Uniformitätsthese und der traditionellen Stellungsfelder einig, was den Wortstatus und die Kategorienindifferenz von C bzw. der linken Satzklammer betrifft.

Für eine kritische Diskussion zu den Konsequenzen dieser Annahme und eine Abwägung gegenüber der sog. Differenzthese, welche unterschiedliche Grundstrukturen für Neben- wie für Hauptsätze annimmt, vergleiche (von Stechow und Sternefeld 1988, 388ff.).

(106) a. [$_{CP}$[$_{C'}$[$_{C}$ weil] [$_{IP}$ diese These nicht stimmen kann]]]
b. [$_{CP}$ [$_{NP}$ diese These]$_j$ [$_{C'}$[$_{C}$ kann$_i$][$_{IP}$ t$_j$ nicht stimmen t$_i$]]]
c.

VF	LSK	MF	RSK
–	weil	diese These nicht	stimmen kann
diese These	kann	nicht	stimmen

Schon in Reis (1985) finden sich verschiedene deutsche Sprachdaten, welche sich nur schlecht mit einem Theorieansatz erklären lassen, der als Füllung für die

Position der Komplementierer phrasale Elemente ausschliesst. Insbesondere Koordinationsdaten, die sie im folgenden Abschnitt gesondert betrachtet werden, führt sie ins Feld.

Im Rahmen der HPSG haben Kathol und Pollard (1995) unter starker Berücksichtigung der Daten von Reis (1985) für ein einziges nebensatzeinleitendes Feld argumentiert (*cf-field*)[42], das nebst Komplementierern auch W-Phrasen und Relativanschlüsse beinhaltet. In Hauptsätzen ist dieses *cf-field* mit dem finiten Verb gefüllt. Das Vorfeld als Landeplatz für Topikalisierung ist für sie im Standard-Deutschen nur in Hauptsätzen verfügbar. Somit vertritt dieser Ansatz die Differenzthese. Um das bei gewissen süddeutschen Dialekten gängige Doppelvorkommen von W-Phrase und Komplementierer wie in (107) zu erklären, erlauben sie in diesen Dialekten auch ein Vorfeld (*vf-field*) in Nebensätzen.

(107) Ich frage mich, [$_{vf}$ wen] [$_{cf}$ dass] [$_{mf}$ Adam] [$_{vc}$ sah]

Dieser Ansatz hat keine grundsätzlichen Probleme damit, dass die Nebensatzeinleitung auch als koordinierte Wortgruppe geschehen kann.

Im Gegensatz zu Kathol, welcher mit dem *cf-field* eine strukturell möglichst einfache Erklärung sucht für die linke Satzperipherie im Nebensatz, geht Bayer (2002) in eine ganz andere Richtung. Im Sinn von Rizzi (1997), welcher den CP/IP-Komplex viel feiner aufgliedern möchte, um grammatische Phänomene wie Finitheit, Topikalisierung oder Fokus phrasenstrukturell repräsentieren zu können, vertritt er für die Kategorie C eine sogenannte C-Split-Hypothese. Darin werden Daten aus süddeutschen Dialekten und dem Niederländischen mit Doppelfüllungen wie in (107) im Vergleich zu typologisch unterschiedlichen Sprachen wie Japanisch und Koreanisch bezüglich Fragekonstruktion dahingehend interpretiert, dass ausserhalb der CP eine sogenannte disjunktive Phrase (disjP) existiert. Diese enthält auf Grund von Bayers fragesemantischen Überlegungen dann die W-Phrase.

In der neuesten, dem minimalistischen Programm verpflichteten Generativen Grammatik für Deutsch von Sternefeld (2005) wird hingegen die Tradition fortgeführt, welche das Vorfeld mit der Spezifikatorposition von C (SpecC) identifiziert und die linke Satzklammer mit der Kopfposition C der CP. Sie stimmt so weitestgehend mit den Annahmen zur Feldfüllung im Grammatik-Duden überein. Sie muss für die Behandlung von koordinierten Komplementiereren jedoch zu einem Ellipsenmodell oder zum Aufeinanderlegen von Syntaxbäumen (3-dimensionales Koordinationsmodell) greifen. Vergleiche dazu den Anhang zum Thema „Koordination“ in Sternefeld (2005).

Insgesamt zeigen sich also in der Sprachwissenschaft zu diesem Thema unterschiedliche und stark wechselnde Auffasssungen.

[42] Nebst den normalen an den deutschen Bezeichnungen orientierten Felderkürzeln „vf“ für Vorfeld und „mf“ für Mittelfeld verwenden Kathol und Pollard (1995) „cf“ für „comp/finite“ und „vc“ für „verb cluster“. Im neueren Aufsatz Kathol (2001) ist die gleiche Argumentation für eine gemeinsame Position von Komplementierern und W-Elementen in Nebensätzen ebenfalls wiedergegeben.

CCP vs. CO Neben der reinen Koordination von KOUS kommt auch die Verknüpfung mit sogenannten Interrogativadverbien (PWAV) wie in (108a) oder mit W-Phrasen mit attribuierenden Interrogativpronomen (PWAT) wie in (108b) vor.

(108) a. Die Beantwortung der Frage, [$_{CO-CP}$ [$_{KOUS}$ ob] [$_{KON}$ und] [$_{PWAV}$ wie]] Kunst "Sinn" macht, [...] [N_{1841}]

b. Immer mehr Ärzte machen sich deshalb Gedanken, [$_{CO-MO}$[$_{KOUS}$ ob] [$_{KON}$ und] [$_{PP}$ auf [$_{PWAT}$ welche] Weise]] dem Brustkrebs vorgebeugt werden könnte. [N_{13857}]

Wie in den vorgehenden Abschnitten diskutiert, werden unterordnende Konjunktionen und W-Elemente sowohl in der Duden-Grammatik wie in der neueren generativen Grammatik strukturell verschieden analysiert. Somit ergibt sich für obige Fälle eine Koordination von Elementen, welche stark abweichenden Status haben. Dafür ist im NEGRA-Annotationsschema die Phrasenkategorie CO reserviert, welche auch konsequent verwendet wird.

Wer für diese Fälle die Hypothese retten will, dass „Koordination nicht über Feldgrenzen hinweg geschehen darf“ (Müller und Ule 2001, 237), muss auf der lexikalischen und/oder syntaktischen Ebene andere Annahmen treffen, als sie in den beiden vorausgehenden Abschnitten geschildert wurden.

Eine einfache lexikalische Möglichkeit für Interrogativadverbien wie in (108b) schlägt (Steiner 2004, 168) in ihrer distributionell begründeten Wortartenklassifikation vor. Unter Referenz auf Bergenholtz und Schaeder (1977) und bewusst im Gegensatz zu Standardwerken wie Eisenberg (1999), werden Interrogativadverbien nur in direkten Fragekontexten als solche kategorisiert. Die (belegten) Beispiele aus (Steiner 2004, 168) sind

(109) a. [$_{PWAV}$ Warum] sind wir nicht orginell?

b. Deutschland hatte seine innenpolitischen Gründe, [$_{KOUS}$ warum] es sich erst jetzt dazu bereit fühlte.

Bei ihren Experimenten zur quantitativen Erschliessung der Distributionseigenschaften von Wortarten mit Hilfe von *k-means*-basiertem partitivem Clustering von Kotexten hat Steiner (2004, 208) Evidenz für die Kategorisierung von Interrogativadverbien als Subjunktionen in Nebensätzen ausgemacht. Ausdrücke wie „warum“ in (109b) sind in einem Cluster eingeordnet worden, das im annotierten Referenzkorpus etwas mehr als 2000 Vorkommen umfasst und wovon gut 50% als unterordnende Konjunktionen ausgewiesen sind.[43] Da bei diesen gut 1000 Vorkommen aber nicht gesagt wird, wie hoch der Anteil der KOUS darin ist, welche traditionell als PWAV annotiert werden, ist die Schlussfolgerung schlecht überprüfbar, „dass die Klassifikation als subordinierende Konjunktion für nebensatzeinleitende Wortformen wie *warum* tatsächlich sinnvoll ist“ (Steiner 2004, 209).

[43] Die genaue Verteilung des zugehörigen Clusters lautet: 1014 unterordnende Konjunktionen, 453 finite Verben, 273 Relativpronomen, 146 nebenordnende Konjunktionen, 73 finite Hilfsverben, 51 finite Modalverben, 6 Interrogativeadverbien, 4 Adverbien, 1 Artikel sowie 1 Präposition.

Zudem beschränken sich die Kotexte dieses Clusters, wie sie selbst bemerkt, auf die 3 Muster `,#er`, `,#sie` sowie `,#es`, d.h. Komma gefolgt von einem Zwischenelement gefolgt von der Wortform „er“, „sie“ oder „es“. Solche Kontexte werden eher in genügend hoher Anzahl vorkommen, um den Schwellwert zu überschreiten und überhaupt in die zu klassifizierenden Kotexte aufgenommen zu werden. Dies macht deutlich, was es heisst, dass beim Kotext-Clustering eben gerade die ähnlichen Kotexte zusammengefasst werden.[44] Die grundsätzliche Problematik von Ansätzen, welche stark auf lokale Distributionseigenschaften setzen, formuliert Stabler (1998, 72) folgendermassen:

> „The problem is that two expressions can be similar in their distributions, but differ in the structural configurations in which they occur and in properties that are satisfied non-locally.“

Das Problem der ungleichen Koordination von KOUS bleibt allerdings bei Sätzen mit attribuierenden W-Elementen (PWAT) wie in (108b) auf der vorherigen Seite bestehen. Eine andere Möglichkeit, dass diese Konstruktionen syntaktisch etwas homogener erscheinen, erhält man, wenn das Wort „ob“ nicht als KOUS aufgefasst wird. Anhand von rein syntaktischen Daten des Standard-Deutschen lässt sich dies nicht besonders gut motivieren. Wenn man Daten aus deutschen Dialekten und verwandten Sprachen mitberücksichtigt, sieht es besser aus.

Im Englischen sieht das Wort „whether“ als Pendant zu „ob“ nicht bloss wie ein W-Element aus, es verhält sich auch syntaktisch bezüglich W-Bewegung anders als „that“, nämlich ähnlich wie W-Phrasen, über die hinaus W-Bewegung schlecht ist.

(110) a. Which way did you say (that) Bill went?
b. I wonder which way Bill went.
c. I wonder whether Bill went East.
d. * Which way do you wonder whether Bill went?

Mit anderen Beispielen begründet Sternefeld (2005) im Abschnitt zur Komplementiererphrase die Idee, dass „ob“ nicht wie „dass“ als Komplementierer funktioniert. In niederländischen Dialekten kann „of“ bzw. „af“ als Pendant zum deutschen „ob“ von einem „dat“ (deutsch „dass“) gefolgt werden wie in Beispiel (111a). Ähnliches ist in einigen deutschen Dialekten mit Sätzen wie in (111b) möglich – vergleiche auch (Dudenredaktion 2005, § 1347).

(111) a. Ik weet niet of (dat) hij komt
ich weiss nicht ob (dass) er kommt
b. Ich weiss nicht, wer (dass) kommt.

[44] Steiner (2004, 128) selbst sagt dies so: „Kotexte zu clustern führt nicht zu einer Disambiguierung der Wortarten, vielmehr werden lediglich die Kotexte geclustert, deren Zwischenelemente ähnlich sind. Gehören diese Zwischenelemente unterschiedlichen (Wortarten-)Klassen an, so werden ebenfalls Kotexte geclustert, die sich hinsichtlich der Verteilung dieser Klassen ähnlich sind.“

Für Sternefeld wäre dann ein Strukturgerüst wie in (112b) für ob-Sätze anzusetzen.

(112) a. Ich weiss nicht, [$_{CP}$ [$_{SpecC}$ wer] [$_{C}$(dass)] kommt].
b. Ich weiss nicht, [$_{CP}$ [$_{SpecC}$ ob] [$_{C}$] er kommt].

Damit sind „wer“ und „ob“ strukturell parallelisiert. Eine Anfrage nach „ob und wer“ in deutschsprachigen Texten bei Google hat Ende September 2005 über 170’000 Treffer ergeben.

(113) a. Dann kann jeder Bürger selbst entscheiden, ob und wer unterrichtet werden soll.[45]
b. Es hing damit vom Zufall ab, ob und wer im Grundbuch eingetragen war.[46]

Es handelt sich in solchen Fällen nicht um symmetrische Koordination, da die Weglassung vom Konjunktor und dem zweiten Konjunkt zu einem ungrammatischen Resultat führt, bzw. zu einer andern Lesart. Die hohe Anzahl Treffer ist erstaunlich, wenn man dem Kriterium in (Kathol 2001, 39) folgt:

> „Crucially, there is no claim here that all cases of conjoined complementizer and wh expression are necessarily grammatical. Such constructions seem to be subject to the requirement that each conjunct is independently licensed in its syntactic relation with the factor constituent. Hence, if the wh-expression is an obligatory complement, the result is usually far less acceptable.“

Als Beispiel dient ihm (114).

(114) * Ich habe erfahren, [daß und wen] er gesehen hat.

Diese Konstruktion mit „daß“ erweist sich mit der Anfrage „dass und wer“[47] bei Google mit etwas über 100 Treffern tatsächlich als selten. Das Argument von Kathol müsste aber genau gleich auch für „ob und wer“ gültig sein, welche ungemein häufiger anzutreffen sind.

2.4.9 CO

Die Kategorie CO für die Annotation von unterschiedlichen Konjunkten (Brants u. a. 1999, 85) ist auf der reinen Wortebene nur 8 Mal in NEGRA vorhanden und umfasst recht heterogene Konstruktionen. Die lexikalischen Füllungen der Konjunkte sind in (115) aufgelistet:

(115) Lexikalische Füllungen aus NEGRA:
„Ortsbeirats- städtischen“ (1), „Schießbuden dergleichen“ (1), „Vorwärts

[45] `http://www.datenschutz-berlin.de/jahresbe/97/doc/spanndsb.htm`
[46] `http://www.spd-hohen-neuendorf.de/pdf/bodenreform.pdf`
[47] Google bezieht standardmässig das Scharf-S „ß“ sowie „ss“ aufeinander.

rückwärts spurten stoppen drehen wenden“ (1), „Was Wann Wo“ (1), „Wenn Aber“ (1), „einzig allein“ (1), „normal Rechtens“ (1), „ob wie“ (1)

Wie in Tabelle 2.36 auf Seite 72 gezeigt, ist die Überschneidung von CO und CAVP beträchtlich, da letztlich nicht geklärt ist, welche Wortarten als äquivalent gelten für eine Koordinationskonstruktion. In Anbetracht der grosszügigen Konjunktfüllung in CAVP ist die CO-Annotation wie in Beispiel (116) wenig sinnvoll.

(116) Immer wieder bemühte er sich, die Überzahlungen [...] als [$_{CO}$ [$_{ADJD}$ normal und [$_{ADV}$ Rechtens[48]]] zu bezeichnen. [N$_{14048}$]

Da sich viele Adjektive adverbial verwenden lassen, sollte die Koordination von Adjektiven und Adverbien in adverbialer Funktion meines Erachtens immer als CAVP annotiert werden.

In Beispiel (117a) verursacht die bezüglich ADV restriktive STTS-Konvention, welche verlangt, dass eine Wortform nur dann mit ADV getaggt werden darf, wenn es dafür „nur adverbial gebrauchte Formen“ (Schiller u. a. 1999, 23) gibt, dass keine CAVP annotiert wird. Dabei handelt es sich bei „einzig und allein“ sogar um eine adverbiale Paarformel, ähnlich wie bei „ein für allemal“, das zusammen in NEGRA und TIGER in 3 verschiedenen Annotationen vorliegt. In (117b) als NP mit eingebetteter PP, in (117c) als Mehrwort-Adverb und in (117c) als CAVP trotz der kategorial unterschiedlichen Konjunkte.

(117) a. [...] die Krise des Wohlfahrtsstaates [$_{CO}$[$_{ADJD}$ einzig] und [$_{ADV}$ allein]] auf dem Feld der Ökonomie bewältigen zu können . [N$_{1372}$]

b. [...] [$_{NP-MO}$ [$_{ART-NK}$ ein] [$_{PP-MNR}$ [$_{APPR}$ für] [$_{ADV-NK}$ allemal]]] [...] [T$_{16977}$]

c. [...] [$_{AVP-MO}$ [$_{ART-AVC}$ ein] [$_{APPR-AVC}$ für] [$_{ADV-AVC}$ allemal]] [...] [T$_{14799}$]

d. [...] [$_{CAVP-MO}$ [$_{ART-CJ}$ ein] [$_{APPR-CD}$ für] [$_{ADV-CJ}$ allemal]] [...] [N$_{9039}$]

Die Annotation von „ob und wie“ als CO ist in Abschnitt 2.4.8 auf Seite 96 ausführlich diskutiert. Eine Fehlannotation mit CO im Zusammenhang mit Morphemkoordination ist in der Fussnote auf Seite 52 behandelt.

2.5 Koordination von Wortgruppen, Phrasen und Sätzen

In diesem Abschnitt diskutiere ich die Phänomene der Koordination, welche grössere Einheiten als ein Wort in den Konjunkten betreffen. Ausgenommen von der Diskussion in diesem Kapitel bleiben dabei die seltenen Kategorien CVZ, CAC und CCP, welche bei der Wortkoordination umfassend besprochen worden sind.

[48] Grossschreibung wie im Original.

In NEGRA und TIGER sind phrasale Koordinationen (P) etwa doppelt so häufig wie die Wortkoordinationen (W), wie die Auflistungen in (118) zeigen:

(118) a. NEGRA: P (6497, 65.4%), W (3444, 34.6%)
b. TIGER: P (12938, 66.9%), W (6413, 33.1%)

2.5.1 Generelles

Insgesamt finden sich 6489 solcher Konstruktionen im NEGRA. Die Tabelle 2.43 auf der nächsten Seite gibt eine Übersicht, wie sich die Koordinationskategorien darauf verteilen. Die koordinierten NP und S teilen sich 80% aller Vorkommen hälftig.

2.5.2 CNP

Wie in Tabelle 2.1 auf Seite 13 gezeigt, sind CNP die überhaupt am häufigsten vorkommende koordinierte Kategorie. Sie sind es auch bei den Wortgruppen.

2.5.2.1 Der Bau der CNP

Der Bau von CNP mit phrasalen Konjunkten ist enorm vielfältig, wenn man die effektiv annotierten NEGRA-Strukturen zugrunde legt. Die insgesamt 2592 Vorkommen von CNP verteilen sich auf 244 Typen. Die Verteilung der Typen mit mindestens 5 Vorkommen ist in Tabelle 2.44 auf Seite 102 zu finden. Zweiteilige symmetrische und syndetische NP-Koordination stellt dabei die Hälfte aller Fälle und ist die wichtigste Kategorie.

Die Vermischung von Konjunkten, welche nur aus einem Wort (terminales Konjunkt) und aus einer Wortgruppe/Phrase (nicht-terminales Konjunkt) bestehen, sind zu vielfältig, um in ihrer kategoriellen Rohform überblickt zu werden. Eine nützliche abstraktere Sicht fasst alle Folgen von Nicht-Terminalen (markiert als N) und Terminalen (markiert als T) nacheinander zusammen. Zusammengestellt in Tabelle 2.45 auf Seite 103 sieht man, dass die reinen Nicht-Terminal-Koordinationen überwiegen. Bei gemischten Koordinationen, welche aus einer terminalen und einer nicht-terminalen Abfolge bestehen, kommen die schwereren, nicht-terminalen Konjunkte 3 Mal häufiger am rechten Rand vor. In der Zusammenstellung sieht man deutlich, dass dieser strukturelle Effekt nicht besteht bei Konjunkten, welche mehr als einen Wechsel von terminal zu nicht-terminal haben.

Die Verteilung bezüglich der Anzahl der Konjunkte in CNP in NEGRA ist in der Tabelle 2.46 auf Seite 103 zu sehen. Die 2-teiligen Koordinationen stellen gut 77% aller Fälle.

2.5.2.2 Die Funktion der CNP

An welchen Stellen finden sich CNP im Satz und innerhalb von kleineren Strukturen? Eine erste Antwort auf diese Frage liefert die Auswertung in Tabelle 2.47.

NEGRA (total 9941)

Kategorie	Anzahl	in %	Konjunkte	Anzahl	in %
CNP	2592	39.9	P	1868	28.8
			WP	724	11.2
CS	2554	39.4	P	2500	38.5
			WP	54	0.8
CVP	532	8.2	P	444	6.8
			WP	88	1.4
CPP	477	7.4	P	472	7.3
			WP	5	0.1
CAP	167	2.6	WP	97	1.5
			P	70	1.1
CO	158	2.4	P	102	1.6
			WP	56	0.9
CAVP	9	0.1	P	6	0.1
			WP	3	0.0

TIGER (total 19351)

Kategorie	Anzahl	in %	Konjunkte	Anzahl	in %
CNP	5047	39.1	P	3616	28.0
			WP	1431	11.1
CS	4702	36.4	P	4700	36.4
			WP	2	0.0
CVP	1237	9.6	P	1187	9.2
			WP	50	0.4
CPP	1051	8.1	P	1049	8.1
			WP	2	0.0
CAP	575	4.5	WP	358	2.8
			P	217	1.7
CO	286	2.2	P	178	1.4
			WP	108	0.8
CAVP	19	0.1	WP	10	0.1
			P	9	0.1

Tabelle 2.43: Verhältnis der wichtigsten koordinierten Phrasen mit phrasalen Konjunkten in NEGRA und TIGER. Legende zu den Kürzeln in der Spalte „Konjunkte“: WP = mindestens ein Konjunkt aus einem Einzelwort. P = kein Konjunkt aus einem Einzelwort

in %	Anzahl	Tochterkonstituenten	kumulativ
48.5	1257	NP KON NP	48
7.7	200	NN KON NP	56
4.9	128	NP NP KON NP	61
4.0	103	NP NP	65
3.7	95	MPN KON MPN	69
2.6	68	NP KON NN	71
2.0	51	NE KON NP	73
1.7	45	NN NN KON NP	75
1.5	38	NP NP NP	77
1.2	31	NP NP NP KON NP	78
1.0	27	KON NP KON NP	79
0.8	22	NE KON MPN	80
0.7	18	MPN MPN KON MPN	80
0.7	17	NE NE KON NP	81
0.5	14	MPN KON NP	82
0.5	14	NN NP	82
0.5	13	MPN MPN	82
0.5	13	NP KON NE	83
0.5	12	NP NP NP NP	84
0.4	11	NN NN NN KON NP	84
0.4	11	NN NP KON NP	84
0.4	10	NP NN KON NN	85
0.4	10	NP NP NP NP KON NP	85
0.3	9	MPN KON NE	85
0.3	9	NP KON MPN	86
0.3	8	NN NN NP	86
0.3	8	NP KON CNP	86
0.3	8	PPER KON NP	87
0.3	7	KON NN KON NP	87
0.3	7	MPN MPN MPN KON MPN	87
0.3	7	NP KON NP KON NP	88
0.3	7	NP NN KON NP	88
0.2	6	CNP KON NP	88
0.2	6	NE NP	88
0.2	6	NN NP KON NN	88
0.2	5	MPN NE	89
0.2	5	NN KON NN KON NP	89
0.2	5	NP NP KON NN	89
0.2	5	PRF KON NP	89
0.2	5	TRUNC KON NP	89

Tabelle 2.44: Die Verteilung der total 2592 lexikalischen CNP-Tochterkonstituenten in NEGRA. Von den total 244 Typen sind nur diejenigen mit mindestens 5 Vorkommen aufgeführt. Das Type-Token-Verhältnis beträgt 1:10.6 .

in %	Anzahl	Konjunktfolgen	kumulativ
72.1	1868	N	72
18.5	480	T N	91
6.2	162	N T	97
1.4	37	T N T	98
1.2	32	N T N	99
0.2	5	T N T N	100
0.2	4	N T N T	100
0.0	1	N T N T N	100
0.0	1	T N T N T	100
0.0	1	T N T N T N	100
0.0	1	T N T N T N T N T	100

Tabelle 2.45: Die Verteilung der Abfolge von Terminal- und Nicht-Terminalkonstituenten in den Tochterkonjunkten der CNP in NEGRA. Legende: T = mindestens eine Terminalkonstituente, N = mindestens eine Nicht-Terminalkonstituente

in %	Anzahl	Anzahl Konjunkte	kumulativ
77.3	2004	2	77
14.5	375	3	92
4.7	123	4	96
1.7	45	5	98
0.7	18	6	99
0.3	9	7	99
0.3	9	8	100
0.1	3	11	100
0.1	3	13	100
0.0	1	1	100
0.0	1	10	100
0.0	1	9	100

Tabelle 2.46: Verteilung der Konjunktanzahl (d.h. ohne Konjunktoren) der total 2592 CNP in NEGRA. Mittelwert = 2.24 , Standardabweichung = 0.73

Anzahl	in %	Funktion	Mutterkategorie	kumulativ
34.3	888	NK	PP	34
23.2	602	SB	S	58
13.0	337	NK	NP	70
6.6	172	OA	S	77
3.6	93	OA	VP	81
3.4	89	—	—	84
3.4	89	APP	NP	88
3.2	84	GR	NP	91
2.5	65	GR	PP	93
1.7	44	PD	S	95
0.8	21	DA	S	96
0.8	20	APP	PP	96
0.7	19	RE	NP	97
0.7	18	CJ	CNP	98

Tabelle 2.47: Verteilung der grammatischen Funktion in Bezug auf die syntaktische Mutterkategorie von CNP in NEGRA (gezeigt werden nur Fälle mit mehr als 10 Vorkommen)

Viele CNP funktionieren als vollständige Nominalphrasen, was sich an ihrer Satzgliedfunktion wie Subjekt (SB), Objekt (OA) oder ihrer Attributfunktion wie postnominaler Genitiv (GR) ablesen lässt. Die Funktion NK stellt auf Grund der flachen Annotation von NP und PP – wie schon im Abschnitt 2.4.2 auf Seite 63 für die Einbettung der Wortkoordinationen besprochen – keine einheitlich zu fassende syntaktische Funktion dar.

Die Einbettung von CNP in NP Die Tabellenzusammenstellung 2.48 auf der nächsten Seite zeigt die unmittelbare linke bzw. rechte Nachbarschaft in Form ihrer Funktion, d.h. die Konfiguration:

$$[_{NP} \ldots [_{?-\textbf{linke SF}} \ldots] [_{CNP} \ldots] [_{?-\textbf{rechte SF}} \ldots] \ldots]$$

Wortgruppenkoordination zeigt im Gegensatz zu den reinen Wortkoordinationen andere strukturelle Eigenschaften. Wortkoordinationen sind auf der linken Seite zu knapp 70% von attributiven/determinativen Elementen begleitet, während Phrasenkoordinationen das nur in gut 30% der Fälle sind. Der Anteil der CNP, welche die NP auf der linken oder rechten Seite abschliessen, ist jeweils um 10% höher.

Die Einbettung der CNP als NK in PP Die Wortgruppenkoordination in PP ist zu fast 90% in Form von maximalen Phrasen annotiert. Wenn man die andern Fälle genauer betrachtet, zeigen sich jedoch viele problematische Annotationen wie in Bespiel (119).

in %	Anzahl	linke SF
31.5	106	NK
27.6	93	—
23.7	80	MO
14.5	49	CM
2.1	7	MNR
0.3	1	GR
0.3	1	GL

in %	Anzahl	rechte SF
65.0	219	—
11.6	39	MNR
8.6	29	GR
7.4	25	APP
2.7	9	RC
2.4	8	NK
1.8	6	PG
0.3	1	OC
0.3	1	MO

Tabelle 2.48: Verteilungen der Funktionen der Schwesterkonstituenten 337 CNP-Phrasenkoordinationen in NK-Funktion in NP. Legende: SF = Schwesterfunktion

in %	Anzahl	linke SF	rechte SF
89.0	790	AC	—
5.4	48	NK	—
2.7	24	AC	MNR
1.1	10	AC	GR
0.5	4	AC	RC
0.3	3	AC	APP
0.3	3	—	AC
0.2	2	NK	MNR
0.1	1	NK	RC
0.1	1	NK	PG
0.1	1	NK	NK
0.1	1	AC	MO

Tabelle 2.49: Verteilung der CNP in NK-Funktion innerhalb von PP aus NEGRA. Legende: SF = Schwesterfunktion

(119) Doch hängt das Einvernehmen stark an der persönlichen Beziehung [$_{PP}$ zwischen [$_{CNP-NK}$ dem EPLF-Führer Issaias Afwerki und dem äthiopischen Präsidenten Meles Zenawi] , [$_{NP-APP}$ zwei Ex-Kampfgefährten gegen Mengistu, [$_{S-RC}$ die sogar entfernt verwandt sind]]] . [N$_{620}$]

2.5.3 CS

Wie in Tabelle 2.1 auf Seite 13 ersichtlich, sind koordinierte Sätze die zweithäufigst vorkommende koordinierte Kategorie. Sie umfassen gemäss Annotationshandbuch von NEGRA sowohl Koordinationen wie Reihungen.

in %	Anzahl	Tochterkonstituenten	kumulativ
52.0	1328	S KON S	52
34.1	870	S S	86
3.9	99	S S KON S	90
2.5	63	S S S	92
1.2	31	S S S KON S	94
1.0	26	S KON VVFIN	95
0.5	12	S CS	95
0.4	9	S KON KON S	96
0.3	8	CS KON S	96
0.3	8	S S S S	96
0.3	7	CS S	96
0.3	7	S KON PTKNEG	97
0.3	7	S KON S KON S	97
0.2	6	S KON CS	97
0.2	6	S KON S S	98
0.2	5	S S S S S	98

Tabelle 2.50: Die Verteilung der lexikalischen Tochterkonstituenten von total 2554 CS in NEGRA. Von den total 63 Typen sind nur diejenigen mit mindestens 5 Vorkommen aufgeführt. Das Type-Token-Verhältnis beträgt 1:40.5 .

Anzahl	in %	Konjunktfolgen	kumulativ
97.9	2500	N	98
1.7	43	N T	100
0.2	6	N T N	100
0.2	5	T N	100

Tabelle 2.51: Die Verteilung der Abfolge von Terminal- und Nicht-Terminalkonstituenten in den Tochterkonjunkten der CS in NEGRA. Legende: T = mindestens eine Terminalkonstituente, N = mindestens eine Nicht-Terminalkonstituente

2.5.3.1 Der Bau der CS

Die in NEGRA annotierten Tochterkonstituenten von CS-Phrasen sind in der Tabelle 2.50 aufgeführt.

2.5.3.2 Funktion von CS

Wie die Tabelle 2.52 auf der nächsten Seite zeigt, bilden knapp 72% aller CS die Top-Phrase in Sätzen. Wenn man im Vergleich dazu die Funktionen von nicht-koordinierten Sätzen in Tabelle 2.53 auf Seite 108 betrachtet, sieht man, dass sich die Verteilung der Funktionen nur geringfügig unterscheidet. Normale Sätze sind

in %	Anzahl	Funktion	kumulativ
73.3	1826	—	73
8.6	214	OC	82
6.5	162	RC	88
4.1	101	MO	92
2.1	52	SB	95
2.1	52	RE	97
1.5	37	CJ	98
0.8	20	RS	99
0.3	7	NK	99
0.3	7	APP	100
0.2	5	MNR	100
0.2	4	CC	100
0.1	3	PD	100

Tabelle 2.52: Verteilung der Funktionen von CS in NEGRA. Top-Phrasen sind mit '—' ausgezeichnet.

nur zu 55% Top-Phrasen, dafür noch zu knapp 19% Konjunkte von CS, was insgesamt ziemlich genau den 74% entspricht, welche die Top-CS und die CS-CJ zusammen ausmachen.

2.5.3.3 Ellipsen in koordinierten Sätzen

Wie in Abschnitt 2.4.6 auf Seite 87 erwähnt, werden elliptische Sätze normalerweise wie vollständige Sätze behandelt. Im Folgenden soll untersucht werden, welche Art von Unvollständigkeit in NEGRA wie häufig ist.

S-Konjunkte ohne finite Verbal-Köpfe Die Tabelle 2.54 auf Seite 109 enthält eine detaillierte Aufstellung über das Vorhandensein einer Kopf-Kategorie (HD) innerhalb der S-Konjunkte von CS. Es wurden ausschliesslich Konfigurationen vom Typ $[_{CS}\ [_{S-CJ}\ [_{?-HD}\ ]]]$ ausgewertet[49]. Nicht einbezogen sind insbesondere auch die Fälle, wo einzelne finite Verbalformen als Konjunkte erscheinen. Über 75% aller CS haben alle S-Konjunkte finite Verbal-Köpfe. Köpfe fehlen typischerweise in den hinteren Konjunkten, d.h. es gibt Rechtsellipsen.

Eine Auswertung, welche beliebig lange Folgen von $\pm$-kopfhaltigen Konjunkten jeweils durch ein $+$ bzw. $-$ zusammenfasst, ergibt folgendes Bild: „+“ (2043, 82.0%), „+ –“ (280, 11.2%), „n/a“ (77, 3.1%), „– +“ (44, 1.8%), „–“ (34, 1.4%), „+ – +“ (11, 0.4%), „– + –“ (1, 0.0%) . Dabei treten die beiden Hauptfälle deutlich hervor. Der Fall „n/a“ wurde kategorisiert, wenn weniger als 2 S-Konjunkte in einer CS vorkommen.

[49]Es gibt einige Annotationsausreisser, wo nicht alle Konjunkte satzartig sind und eigentlich als CO annotiert sein müssten.

in %	Anzahl	Funktion	kumulativ
55.1	15980	—	55
18.6	5385	CJ	74
7.9	2291	RC	82
6.6	1908	OC	88
5.4	1572	MO	94
2.1	619	RE	96
1.5	427	SB	97
0.6	184	RS	98
0.5	156	MNR	98
0.5	154	APP	99
0.4	114	CC	99
0.4	106	DH	100
0.2	45	NK	100
0.1	30	PD	100
0.0	3	DM	100
0.0	2	OA	100

Tabelle 2.53: Verteilung der von S in NEGRA. Top-Phrasen sind mit '—' ausgezeichnet. Es sind nur Funktionen aufgeführt, welche mindestens 2 Mal vorkommen.

Der Satz (120a) ist typisch für Verb-Auslassung im 2. Konjunkt, wie er unter dem Etikett „+ –“ zusammengefasst wird. Ähnlich im Satz (120b), wo allerdings das fehlende Hilfsverb im 2. Satzkonjunkt im Plural („wurden“) und nicht im Singular wie im 1. Konjunkt erscheinen muss.

(120) a. [$_{CS}$ [$_{S}$ Ein Wort-Redakteur sorgt für kurze, eingespielte Beiträge] , [$_{S}$ ein Musikredakteur für die richtigen Platten]] . [N_{296}]

b. [$_{CS}$ [$_{S}$ Jetzt wurde eine " königlich-preußische Polizeiverwaltung " errichtet] , [$_{S}$ dem ersten Polizeipräsidenten Guido von Madai alle Polizeibefugnisse übertragen]] . [N_{172}]

Ein Beispiel für die selteneren Konstruktionen vom Typ „– +“ ist im Relativsatz von Beispiel (121a) zu finden, mit einer typischen Kontrastmarkierung mittels dem paarigen „nur. . . hingegen“. Es gibt unter dem Kürzel „– +“ auch Fälle wie in (121b), wo eine Absolutkonstruktion vorliegt und keinerlei syntaktische Parallele zwischen den beiden Sätzen herrscht. Dieser in der traditionellen Grammatik bekannte Konstruktionstyp erhält leider in NEGRA keine eigene syntaktische Funktion zugeschrieben und lässt sich aufgrund der annotierten Struktur nicht unterscheiden. Eine qualitative intellektuelle Evaluation ergibt, dass von allen Fällen des Typs „– +“ mit 25 Fällen etwas mehr als die Hälfte echte Linksellipsen sind.

(121) a. [. . .] [$_{CS-RC}$ [$_{S}$ auf der er nur als "Führer" der türkischen Zyprer bezeich-

in %	Anzahl	± kopfhaltig	kumulativ
75.1	1916	+ +	75
9.1	233	+ –	84
4.9	126	+ + +	89
3.4	87	n/a	92
1.8	45	– +	94
1.2	31	– –	96
1.1	29	+ + + +	97
0.9	24	+ – –	98
0.6	16	+ + –	98
0.4	10	+ + + + +	98
0.2	6	+ – +	99
0.2	5	+ – – –	99
0.1	3	– – – –	99
0.1	3	– + +	99

Tabelle 2.54: Übersicht zur Annotation von HD-Funktionen bei Konjunkten vom Typ S in NEGRA. Von den 2551 Fällen werden in der Tabelle nur solche mit einem Mindestvorkommen von 3 angezeigt. Insgesamt gibt es 26 verschiedene Typen. Lesebeispiel für die Spalte „±-kopfhaltig“: „+ – –“ bedeutet, dass in einer CS-Phrase, welche aus 3 S-Konjunkten besteht, das 1. Konjunkt einen finiten Verbal-Kopf besitzt, das 2. und 3. Konjunkt jedoch keinen finiten Verbal-Kopf haben. Alle Konjunkte, welche nicht vom Typ S sind, sind nicht sichtbar. Falls nicht mindestens 2 S-Konjunkte erscheinen in einer CS-Phrase, wird der Wert „n/a“ vergeben, da hier nur das Vorkommen von Verbal-Köpfen in mehreren S-Konjunkten interessiert.

net] , [$_S$ sein Kollege Vassiliou hingegen als Präsident der griechischen Zyprer aufgeführt war]] . [N_{683}]

b. [$_{CS}$ [$_S$ Kein Sommer ohne den krächzenden Barden in Deutschland] , [$_S$ bei irgendeinem Freiluftkonzert landauf-landab ist er garantiert mit von der Partie]] . [N_{8767}]

Vollständig kopflose Satzreihungen, welche unter dem Kürzel „−“ aufgeführt werden, sind mit Beispiel (122a) illustriert und dienen als wirkungsvolles Stilmittel zur lebendigen Verdichtung und Intensivierung der Äusserungen. Das Anhängen von kopfhaltigen Konjunkten an eine Linksellipse (Etikett „+ -- +“) tritt sporadisch auf wie in Beispiel (122b). Hier muss man sich fragen, ob ein Zusammenfassen des 1. Konjunkts mit seiner rechtselliptischen Erweiterung strukturell nicht sinnvoller wäre.

(122) a. [$_{CS}$ [$_S$ Haydn somit kompakt] , [$_S$ Haydn dramatisch in seiner "ungeschminkten", unmißverständlichen Kristallisation]] . [N_{17214}]

b. [$_{CS}$ [$_S$ 30 Parlamentarier starben in diesen zwei Dekaden] , [$_S$ manche eines natürlichen Todes] , [$_S$ andere wurden ermordet]] . [N_{17214}]

Subjektlose S-Konjunkte mit finiten Verbal-Köpfen Die Tabelle 2.55 auf der nächsten Seite gibt eine detaillierte Aufstellung über das Vorhandensein von Subjekten (SB) und finiten Verbalköpfen (V?FIN) innerhalb der Konjunkte von CS. Es wurden ausschliesslich Konfigurationen vom Typ [$_{CS}$ [$_{S-CJ}$ [$_{?-SB}$ ...]]] sowie [$_{CS}$ [$_{V?FIN-CJ}$]] ausgewertet. Die Kreuzklassifikation von ±-subjekthaltig und ±-verbalkopfhaltig ist folgendermassen kodiert:

	+subjekt	-subjekt
+verbal	+	>
-verbal	<	–

Diese Einteilung in Subtypen ergibt in NEGRA mit 62 Typen eine recht grosse Anzahl von verschiedenen Konjunktfolgen.

Dabei treten die beiden recht ausgewogenen Hauptfälle „+“ sowie „+ >“ mit knapp 75% noch deutlicher hervor. 2-teilige Koordinationen mit Subjektlücken (SLK) im 2. Konjunkt sind mit fast 34% eine erstaunlich häufige Konstruktion[50]. In Beispiel (123a) ist eine typische syndetische SLK-Koordination illustriert, aber auch asyndetische SLK wie in Beispiel (123b) sind häufig. Eine 3-teilige monosyndetische vom Typ „+ > >“ illustriert (123c).

(123) a. [$_{CS}$ [$_S$ Ghozali stuft den Öffnungsprozeß von 1988 bis 1991 als einen Unfall ein] und [$_S$ will ihn " ungeschehen " machen]] . [N_{3136}]

b. [$_{CS}$ Sie gehen gewagte Verbindungen und Risiken ein] , [$_S$ versuchen ihre Möglichkeiten auszureizen]] . [N_2]

[50]Eine Spezialität des Gojol-Parsers aus Abschnitt 3.3 auf Seite 176 ist es, dass er solche Konstruktionen recht gut erkennt.

in %	Anzahl	Konjunkttypen	kumulativ
39.6	1010	+ +	40
33.9	866	+ >	74
5.2	133	+ –	79
4.8	122	+ <	84
2.2	55	n/a	86
2.2	55	+ > >	88
1.6	42	+ + +	90
1.5	37	< +	91
0.9	24	> +	92
0.7	19	+ < <	93
0.6	16	– –	93
0.6	15	< <	94
0.6	15	+ + >	94
0.5	12	+ > > >	95
0.5	12	+ > +	95
0.4	10	> >	96
0.3	8	+ + + +	96
0.2	6	+ < +	96
0.2	6	+ + <	96
0.2	5	> <	97
0.2	5	+ > –	97
0.2	5	+ + –	97
0.2	4	> –	97
0.2	4	– +	98
0.2	4	+ > > > >	98
0.2	4	+ < < <	98
0.2	4	+ – >	98
0.2	4	+ – –	98
0.2	4	+ + + + +	98
0.1	3	< >	99
0.1	3	< < < <	99
0.1	3	< + +	99

Tabelle 2.55: Verteilung der Annotation von SB- und HD-Funktionen bei Konjunkten vom Typ S in NEGRA. Angezeigt werden nur Fälle mit mindestens 3 Vorkommen. Insgesamt gibt es 62 verschiedene Typen. Falls nicht mindestens 2 interessierende Konjunkte erscheinen in einer CS-Phrase, wird der Wert „n/a“ vergeben, da hier nur der Vergleich von Vorkommen in mindestens 2 Konjunkten interessiert. Lesebeispiel für „+>>>“: Ein einziges S-Konjunkte, welches sowohl einen finiten Verbalkopf als auch ein Subjekt aufweist, wird gefolgt von 3 Konjunkten, bei denen ein finiter Verbalkopf, aber kein Subjekt vorkommt.

c. [$_{CS}$ [$_{S}$ Am letzten Hindernis blieb sein Pferd hängen], [$_{S}$ überschlug sich] und [$_{CS}$ erdrückte den Reiter]] . [N_{757}]

Der Subtyp „> +“, bei dem die Subjektlücke im 1. Konjunkt auftritt, soll noch genauer betrachtet werden. In (124a) auf dieser Seite fehlt dem 1. Konjunkt semantisch nicht nur das Subjekt, sondern der Satz enthält eigentlich 2 koordinierte finite Verben, welche zu allen restlichen Satzgliedern in derselben Funktion stehen. Auf der primären Strukturebene wird dies allerdings in keiner Form annotiert, ebensowenig wie in (124b).

(124) a. [$_{CS}$ [$_{S}$ Als rohstoffarmes Land war] und [$_{S}$ ist Dänemark ökonomisch auf die Ausbildung und Entfaltung des schöpferischen Potentials seiner Gesellschaft angewiesen]] . [N_{1229}]

b. [$_{CS}$ [$_{S}$ Seit 20 Jahren organisiert] und [$_{S}$ moderiert Oertl Veranstaltungen des Frankfurter Blindenbundes [. . .]]] . [N_{1526}]

Die relativ hohe Zahl von Subtypen in Tabelle 2.55 auf der vorherigen Seite lässt sich mit der verdichtenden Darstellung wie in Tabelle 2.56 auf der nächsten Seite, wo Konjunktfolgen mit identischen Subtypen nur noch durch ein Symbol repräsentiert werden, auf 27 Typen reduzieren.

SLK und sekundäre Kanten in TIGER Die mangelhafte Annotation der Argumentstruktur in koordinierten Sätzen sollte ursprünglich schon in NEGRA mithilfe von sekundären Kanten behoben werden (Skut u. a. 1997, 91): „Structure sharing is expressed using secondary links“. Dies wurde allerdings erst im TIGER-Korpus realisiert. Die Abbildung 2.10 auf Seite 114 illustriert, wie das explizite Subjekt des 1. Konjunkts als sekundäre Kante an den S-Knoten des 2. Konjunkte angehängt wird.[51]

In Tabelle 2.57 auf Seite 115 ist die Verteilung der subjektlosen, aber verbalkopf enthaltenden CS-Koordination aus dem TIGER-Korpus zusammengestellt, welche der NEGRA-Tabelle 2.55 auf der vorherigen Seite entspricht. Im Gegensatz zu NEGRA gibt es im TIGER-Korpus sogar etwas mehr CS mit Konjunkten vom Typ „+ >“ als „+ +“.

Mit Hilfe der annotierten sekundären Kanten in TIGER lassen sich nun die echten Subjekt-Ellipsen klar zuverlässiger erkennen. Die Tabelle 2.58 auf Seite 116 enthält die Verteilung der Konjunkte mit sekundär ergänzten Subjekten, welche in der Tabelle mit „>/SB“ markiert sind. Bei den 2-teiligen Koordination stehen den knapp 73% „+ >/SB“ nur gut 8% „+ >“ gegenüber. Auf NEGRA übertragen sollten also auch etwa 9/10 der „+ >“-Typen SLK sein. Bei den drei-teiligen Konjunkten sind in TIGER die Verhältnisse bezüglich „+ > >“ : „+ >/SB >/SB“ mit ca. 1:25 noch deutlicher.

[51] Ein ärgerlicher Programmfehler im TIGERSearch-Werkzeug bis und mit Version 2.1.1 verhindert, dass mehr als eine sekundäre Kante ausgehend vom gleichen Knoten mit der gleichen Funktionsbeschriftung zu verschiedenen Zielknoten angezeigt wird, obwohl diese Kanten durchaus im Korpus annotiert sind.

in %	Anzahl	Konjunktfolgetypen	kumulativ
41.7	1064	+	42
37.5	957	+ >	79
6.1	155	+ <	85
5.6	144	+ –	91
2.2	55	n/a	93
1.6	42	< +	95
1.0	25	> +	96
0.7	19	<	96
0.7	17	–	97
0.5	14	+ > +	98
0.4	11	>	98
0.4	9	+ < +	98
0.2	5	> <	99
0.2	5	+ > –	99
0.2	5	+ – >	99
0.2	4	> –	99
0.2	4	– +	99
0.1	3	< >	100
0.1	3	+ > + >	100

Tabelle 2.56: Verteilung von SB- und HD-Funktionen bei Konjunkten vom Typ S und V?FIN in NEGRA mit Verdichtung von Konjunktfolgen zu identischen Subtypfolgen. Insgesamt gibt es 27 Typen. Dargestellt in der Tabelle sind nur die 19 Fälle mit mindestens 3 Vorkommen.
Lesebeispiel zu „+>“: Eine beliebige Folge von S-Konjunkten, welche sowohl einen finiten Verbalkopf als auch ein Subjekt aufweisen, werden gefolgt von mindestens einem Konjunkt, bei dem ein finiter Verbalkopf, aber kein Subjekt vorkommt.

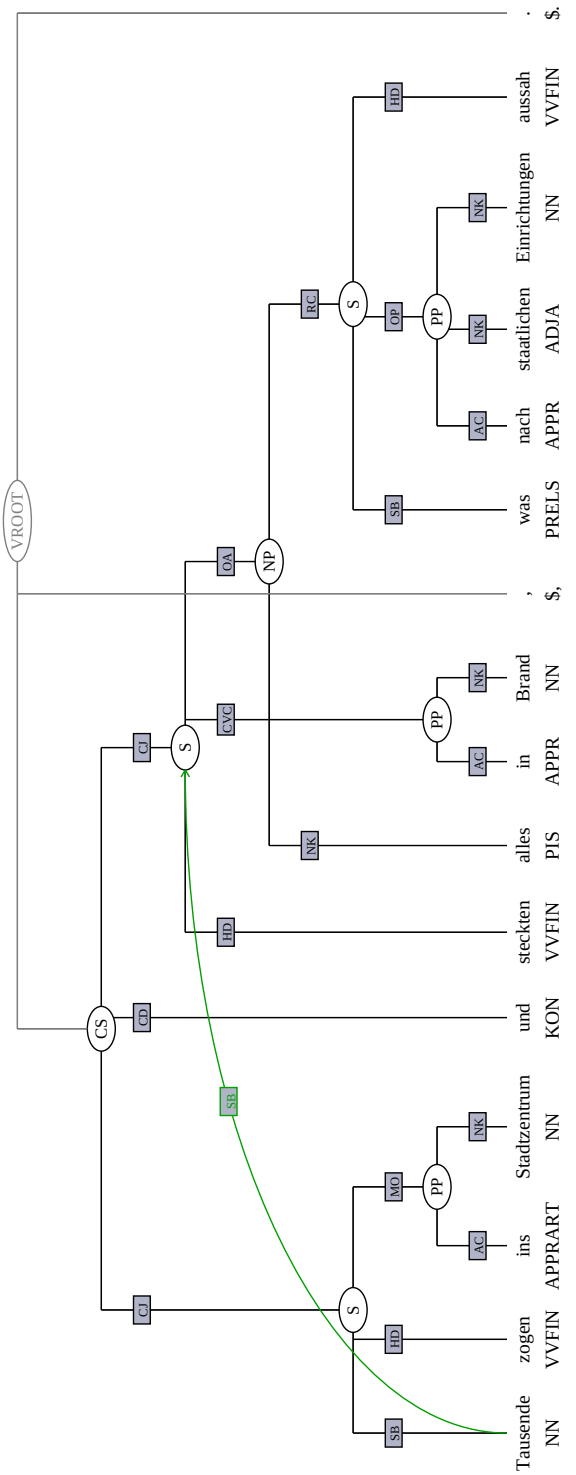


Abbildung 2.10: Satz 355 mit Subjekt-Lücken-Konstruktion aus TIGER als Baumdiagramm mit sekundärer SB-Kante

in %	Anzahl	Konjunkttypen	kumulativ
36.5	1713	+ >	36
35.1	1649	+ +	72
6.6	308	+ <	78
3.7	172	+ –	82
2.7	128	+ > >	85
2.2	102	+ + +	87
2.0	94	> +	89
1.7	78	n/a	90
1.5	71	< +	92
0.8	38	– –	93
0.7	31	< <	94
0.6	28	– +	94
0.5	25	+ + >	95
0.5	24	+ > +	95
0.5	22	< >	96
0.4	21	> >	96
0.4	17	+ < <	96
0.3	16	+ > > >	97
0.3	12	+ + + +	97
0.2	11	> –	97
0.1	7	< < <	97
0.1	7	+ + <	97
0.1	6	< < +	98
0.1	6	+ > –	98
0.1	6	+ < +	98
0.1	5	– – –	98
0.1	5	+ > > > >	98
0.1	5	+ – –	98
0.1	5	+ + + + +	98

Tabelle 2.57: Verteilung der Annotation von SB- und HD-Funktionen bei Konjunkten vom Typ S in TIGER. Angezeigt werden nur Fälle mit mindestens 5 Vorkommen. Insgesamt gibt es 78 verschiedene Typen. Falls nicht mindestens 2 interessierende Konjunkte erscheinen in einer CS-Phrase, wird der Wert „n/a“ vergeben, da hier nur der Vergleich von Vorkommen in mindestens 2 Konjunkten interessiert.

in %	Anzahl	Konjunkttypen	kumulativ
72.9	1483	+ >/SB	73
8.4	171	+ >	81
5.7	116	+ >/SB >/SB	87
3.6	74	> +	91
0.9	19	+ >/SB +	92
0.9	18	>/SB +	92
0.8	17	+ + >/SB	93
0.7	15	< >/SB	94
0.7	15	+ >/SB >/SB >/SB	95
0.7	14	> >	95
0.4	9	> –	96
0.3	6	+ >/SB –	96
0.3	6	+ + >	96
0.2	5	+ > >	96
0.2	5	+ > +	97
0.2	5	+ >/SB >/SB >/SB >/SB	97
0.2	4	> > >	97
0.2	4	> <	97
0.2	4	< >	98

Tabelle 2.58: Verteilung von Subjekt-Lücken-Konstruktionen in TIGER. Kontrastiert werden alle Konjunkttypen mit mindestens einem Konjunkt vom „>“-Typ, d.h. wo ein finiter Verbalkopf ohne Subjekt erscheint. Die Zeichenfolge „>/SB“ bedeutet, dass eine sekundäre SB-Kante an das Konjunkt annotiert ist. Insgesamt gibt es 52 Typen. Dargestellt in der Tabelle sind nur die Typen mit mindestens 4 Vorkommen.
Lesebeispiel zu „+ >/SB“: Mindestens ein satzwertiges S-Konjunkt ist gefolgt von einem Konjunkt mit finitem Verbkopf, aber ohne Subjekt, bei dem jedoch eine sekundäre Kante das Subjekt des 1. Konjunkt im subjektlosen Konjunkt ergänzt. Dies entspricht der Struktur in Abbildung 2.10 auf Seite 114.

in %	Anzahl	in %	Kopfdistanz	Anzahl	in %
+ >/SB	1483	98.8	10+	444	29.6
			6	183	12.2
			5	180	12.0
			7	145	9.7
			4	134	8.9
			8	127	8.5
			9	119	7.9
			3	95	6.3
			2	56	3.7
>/SB +	18	1.2	2	18	1.2

Tabelle 2.59: Verteilung der Tokendistanz der Verbalköpfe in 2-teiligen Konjunkten nur der Typen „+ >/SB“ und „>/SB +“ mit elliptischem Subjekt in TIGER. Mittelwerte der Distanz von „+ >/SB“ ist 8.36 bei einer Standardabweichung von 5.02 .
Lesebeispiel für 1. Zeile: In 444 Fällen sind bei CS-Kordinationen mit 2 Konjunkten des Typs „+ >/SB“ die finiten Verbalköpfe der beiden Konjunkte mehr als 10 Token voneinander entfernt.

Es gibt in der Tabelle 2.58 auf der vorherigen Seite einige Spezialfälle. Die Vorkommen von Rechtsellipsen mit dem Subtyp „>/SB+“ sind wie in den Beispielen (124) auf Seite 112 durchwegs eng koordinierte finite Verben, wie sie auch in Beispiel (125) in Verbzweitstellung mit dem Subjekt im Mittelfeld erscheint. Eine partielle syntaktische Analyse hätte es in diesem Beispiel sehr schwierig zu erkennen, dass „starb“ nicht in die Reihe von „sang“ und „komponierte“ gehört.

(125) [$_{CS}$ [$_{S}$ Wenige Wochen bevor der Queen-Sänger [...] starb , sang] und [$_{S}$ komponierte er in seinem Studio in Montreux noch mehrere neue Songs]] . [T$_{5800}$]

Um zu überblicken, welche Konstruktionen echte SLK darstellen und welche Konstruktionen eher eng koordinierte finite Verben sind, wurde die Distanz der finiten Verbalköpfe in Konjunkten des Typs „>/SB +“ und „+ >/SB“ untersucht. Wie in der Tabelle 2.59 ersichtlich, ist bei den Konjunkten vom Typ „>/SB +“ in allen Fällen genau ein Konjunktor zwischen den finiten Verbalköpfen. Während bei den Konjunkten vom Typ „+ >/SB“ ein Abstand im Mittelwert von 8.36 besteht und nur wenige finite Verbalköpfe eine Tokendistanz von 2 haben.

2.5.4 CPP

Wie in Tabelle 2.1 auf Seite 13 ersichtlich, sind CPP mit 477 Vorkommen in NEGRA die fünfthäufigste Koordinationskategorie. Fast alle Fälle bestehen nur aus phrasalen Konjunkten, nur eine Handvoll enthält mindestens ein Einzelwortkon-

in %	Anzahl	Tochterkonstituenten	kumulativ
67.7	323	PP KON PP	68
12.8	61	PP PP	80
5.7	27	PP PP KON PP	86
4.6	22	PP PP PP	91
1.9	9	KON PP KON PP	93
1.5	7	PP KON PP KON PP	94
1.3	6	PP PP PP KON PP	96
1.0	5	PP PP PP PP	96
0.4	2	CPP KON CPP	97
0.4	2	PP KON CPP	97
0.4	2	PP PP PP PP PP	98

Tabelle 2.60: Die Verteilung der lexikalischen Tochterkonstituenten von total 477 CPP in NEGRA. Von den 22 Typen sind nur diejenigen mit mindestens 2 Vorkommen aufgeführt. Das Type-Token-Verhältnis beträgt 1:21.7.

junkt, wobei sich diese Fälle bei genauerer Betrachtung fast allesamt als problematisch (Konjunkt „usw.“) oder Annotationsfehler erweisen.

2.5.4.1 Der Bau der CPP

Die häufigsten der 22 verschiedenen Typen der in NEGRA annotierten Tochterkonstituenten von CPP-Phrasen sind in Tabelle 2.60 aufgeführt. Nominalphrasen und Präpositionalphrasen sind sich als Konstituenten strukturell am ähnlichsten. Wenn man die Verteilung der Koordinationsmittel bei CNP vergleicht mit den Koordinationsmitteln in CPP, zeigen sich deutliche Unterschiede:

(126) a. Verteilung der Koordinationsmittel in CNP in NEGRA:
„syn“ (3959, 76.6%), „mono“ (742, 14.4%), „asyn“ (422, 8.2%), „x“ (46, 0.9%)

b. Verteilung der Koordinationsmittel in CPP in NEGRA:
„syn“ (345, 72.3%), „asyn“ (94, 19.7%), „mono“ (37, 7.8%), „x“ (1, 0.2%)

Asyndetische Koordinationen sind bei CPP mit fast 20% vertreten. Das Beispiel (127a) ist illustrativ für Reihungen, wo im 2. Konjunkt noch zusätzlich die Präposition ausgespart wird. Beispiel (127b) zeigt eine Reihung, wo dieselbe Präposition dreimal verwendet wird.

(127) a. Zugleich hat Genazino sein Augenmerk auch immer wieder (wie in einer Serie in der FR-Wochenendbeilage) [$_{CPP}$ [$_{PP}$ auf das Abbild] , [$_{PP}$ auf Fotografien oder Malerei]] , gerichtet. [N$_{137}$]

b. Die übrigen siebzehn bis zwanzig Wochen nützten sie [$_{CPP}$ für selbständiges Studium, für Lektüre, für die Vorbereitung der Lehrveranstaltungen des folgenden Semesters] . [N$_{13722}$]

Weiter gibt es einige Fälle wie in (128), wo die Tochterkonjunkte nicht durch Komma getrennt sind. Solche „Koordinationen“ treten gerne mit Präpositionsfüllungen wie „von. . . bis“ auf und drücken meist eine Erstreckung aus. Eine koordinative Annotation wird in NEGRA eigentlich nur dann gemacht, wenn die topologischen Verhältnisse keine Folge von PP erlauben an dieser Stelle. Das kann im Vorfeld sein wie in (128a) oder in Titeln, welche nur aus einer Phrase bestehen wie in (128b). Auch für 3-teilige Erstreckungsausdrücke finden sich Beispiele in NEGRA, welche gerne mit „von. . . über. . . bis“ gebildet werden wie in der Apposition in (128c).

(128) a. [$_{CPP}$ [$_{PP}$ Von Januar] [$_{PP}$ bis Mai] dieses Jahres] lief im Außenhandel ein gegenüber dem entsprechenden 91er Zeitraum rund verdoppelter Überschuß von 10,7 Milliarden Mark auf. [N$_{5999}$]

b. Lothar Kölm (Herausgeber): Parteichefs im Kreml. [$_{CPP}$ [$_{PP}$ Von Lenin] bis [$_{PP}$ Gorbatschow]] . [. . .] [N$_{2536}$]

c. [. . .] , auch nicht Ergebnis etwaiger Prunkbauten der Wallmann-Ära, [$_{CPP}$ [$_{PP}$ von Alter Oper] [$_{PP}$ über Museumsufer] [$_{PP}$ bis Römerbergbebauung]] , sondern [. . .] [N$_{3798}$]

Rekursiv eingebettete CPP-Konjunkte werden selten annotiert in NEGRA und beinhalten meist Probleme. Anlass sind oft syndetisch koordinierte Erstreckungskombinationen mit „von. . . bis“ wie in (129a), wo kurzerhand die Ortsbestimmung ebenfalls in die CPP-Konjunkte integriert wurde. In (129b) wird die Kardinalzahl „14“ als elliptische PP annotiert, allerdings fehlt sowohl die Präposition „von“ als auch „Uhr“, wobei der Ausdruck „Uhr“ auch im 1. Konjunkt nicht erscheint. Obwohl solche Zeitangaben linguistisch uninteressant sein mögen, scheint ihre Komplexität bezüglich elliptischer Auslassungen umgekehrt proportional zu sein. Solche Zeitausdrücke stellen oft formelhafte Wendungen dar, deren inneren Verhältnisse nur schwierig in ihren syntaktischen Abhängigkeiten bestimmt werden können.

(129) a. Ausleihe und Rückgabe von Lesestoff ist [$_{CPP}$ [$_{CPP}$ [$_{PP-CJ}$ in Walldorf] [$_{PP-CJ}$ vom 22. Juni] [$_{PP-CJ}$ bis zum 12. Juli]] und [$_{CPP}$ [$_{PP-CJ}$ in Mörfelden] [$_{PP-CJ}$ vom 13. Juli] [$_{PP-CJ}$ bis zum 2. August]]] zu den üblichen Zeiten möglich. [N$_{2876}$]

b. Der Schloßherr begrüßt seine Gäste zwischen dem 1. April und 30. November täglich [$_{CPP}$ [$_{CPP}$ [$_{PP-CJ}$ von 10] [$_{PP-CJ}$ bis 12]] und [$_{CPP}$ [$_{CARD-CJ}$ 14 [$_{PP-CJ}$ bis 18 Uhr]]]] [. . .] [N$_{12542}$]

Präpositionen in den Konjunkten von CPP Wie gleichförmig sind die Präpositionen, welche man innerhalb von koordinierten PP findet? Die Auflistung (130)

enthält für NEGRA jeweils die Präpositionenfolge aller 62 Types, welche mindestens 2 Vorkommen haben und wo als Tochterkonjunkt rekursiv keine CPP eingebettet ist.

(130) Verteilung der Präpositionen der Tochter-PP in CPP in NEGRA mit mindestens 2 Vorkommen:
„in in“ (29), „für für“ (23), „von von“ (18), „in im“ (15), „aus aus“ (14), „auf auf“ (14), „mit mit“ (12), „im im“ (11), „um um“ (9), „durch durch“ (9), „an an“ (9), „am am“ (9), „bei bei“ (8), „als als“ (8), „von bis“ (7), „zur zur“ (6), „nach nach“ (6), „in mit“ (6), „zum zur“ (5), „von vom“ (5), „in an“ (5), „gegen gegen“ (5), „zum zum“ (4), „wegen wegen“ (4), „vom vom“ (4), „mit mit mit“ (4), „in auf“ (4), „im in“ (4), „über über“ (3), „zwischen zwischen“ (3), „zu zu“ (3), „zur zum“ (3), „von über zum“ (3), „von zur“ (3), „ohne mit“ (3), „für für für“ (3), „beim bei“ (3), „auf in“ (3), „zu zur“ (2), „zu zum“ (2), „zur zu“ (2), „zum zum zum“ (2), „von für“ (2), „vom zum“ (2), „um um um“ (2), „trotz trotz“ (2), „ohne ohne“ (2), „in vom zum“ (2), „im über“ (2), „im im in“ (2), „im auf“ (2), „für gegen“ (2), „bei im“ (2), „bei bei bei“ (2), „bei beim“ (2), „beim beim“ (2), „aus wegen“ (2), „aus von“ (2), „an in“ (2), „an am“ (2), „am in“ (2), „am an“ (2)

Von den total 1020 Fällen in TIGER ergeben die 90 Types mit mindestens 2 Vorkommen folgende ähnliche Verteilung:

(131) Verteilung der Präpositionen der Tochter-PP in CPP in TIGER:
„in in“ (89), „für für“ (64), „als als“ (55), „von von“ (49), „mit mit“ (48), „auf auf“ (39), „um um“ (38), „bei bei“ (24), „über über“ (23), „in im“ (23), „an an“ (22), „aus aus“ (21), „durch durch“ (19), „im in“ (17), „im im“ (15), „gegen gegen“ (15), „zu zu“ (13), „in bei“ (13), „nach nach“ (11), „zur zur“ (10), „in auf“ (10), „ohne ohne“ (8), „auf in“ (8), „zum zum“ (7), „vor vor“ (7), „im bei“ (7), „am an“ (7), „zwischen zwischen“ (6), „zum zu“ (6), „vom vom“ (6), „in mit“ (6), „gegen für“ (6), „an am“ (6), „mit mit mit“ (5), „im auf“ (5), „für für für“ (5), „am am“ (5), „zur zum“ (4), „von vom“ (4), „von bis“ (4), „nach in“ (4), „in in in“ (4), „bis bis“ (4), „bei in“ (4), „beim bei“ (4), „zu zur zur“ (3), „zu zur“ (3), „zum zur“ (3), „wegen wegen“ (3), „wegen aus“ (3), „von zu“ (3), „unter in“ (3), „um um um“ (3), „in unter“ (3), „in für“ (3), „bei beim“ (3), „beim beim“ (3), „auf im“ (3), „an im“ (3), „an an an“ (3), „über über über“ (2), „über in“ (2), „zur zu zur“ (2), „zur zur zu“ (2), „zum für“ (2), „von über zu“ (2), „von von von“ (2), „von durch“ (2), „von bei“ (2), „von aus“ (2), „von auf“ (2), „vom von“ (2), „ohne mit“ (2), „in in in in“ (2), „in ins“ (2), „in im in“ (2), „in durch“ (2), „in beim“ (2), „im in bei“ (2), „im im im“ (2), „für zur“ (2), „für gegen“ (2), „durch mit“ (2), „durch in“ (2), „auf mit“ (2), „an in“ (2), „an auf“ (2), „am nach“ (2), „als zur“ (2), „als als als“ (2)

In den meisten Fälle sind die Präposition tatsächlich identisch oder Varianten mit oder ohne verschmolzenem Artikel wie etwa „in im“. Ausnahmen bilden die Erstreckungsausdrücke, lokale Präpositionen sowie einige adversative Kombinationen wie „ohne mit“ oder „für gegen“.

Wenn die verschmolzenen Präpositionen vom Typ (APPRART) zu ihren unverschmolzenen Äquivalenten normalisiert werden, vereinheitlicht sich das Bild deutlich.

In Auflistung (132) sind alle normalisierten Präpositionsfolgen aufgeführt aus NEGRA und TIGER, wobei alle adjazenten Folgen identischer Präpositionen nur durch eine Erwähnung ausgedrückt werden und total noch 194 Types bei total 1476 Fällen ergeben. Der Ausdruck „in (229)“ steht in der Auflistung (132) für beliebig viele hintereinander folgende Präpositionen „in“, was insgesamt 229 Mal vorkommt. Der Ausdruck „bei in“ steht für alle Fälle, wo mindestens ein „bei“ von mindestens einem „in“ gefolgt wird.

(132) „in“ (229), „für“ (97), „von“ (92), „zu“ (91), „mit“ (70), „an“ (69), „als“ (67), „bei“ (56), „auf“ (55), „um“ (53), „aus“ (38), „in bei“ (31), „durch“ (30), „über“ (30), „in auf“ (24), „gegen“ (21), „nach“ (18), „auf in“ (16), „an in“ (14), „in mit“ (14), „in an“ (12), „von zu“ (12), „ohne“ (11), „von bis“ (11), „zwischen“ (10), „bei in“ (9), „gegen für“ (8), „von über zu“ (8), „in zu“ (7), „vor“ (7), „wegen“ (7), „bis“ (5), „für zu“ (5), „ohne mit“ (5), „an auf“ (4), „auf mit“ (4), „aus von“ (4), „für gegen“ (4), „in unter“ (4), „nach in“ (4), „unter in“ (4), „von durch“ (4), „von in“ (4), „zu in“ (4), „an bei“ (3), „an nach“ (3), „auf in auf“ (3), „bei zu“ (3), „in als“ (3), „in auf in“ (3), „in für“ (3), „in nach“ (3), „in ohne“ (3), „trotz“ (3), „von für“ (3), „wegen aus“ (3), „als auf“ (2), „als in“ (2), „als zu“ (2), „an zu“ (2), „auf an“ (2), „auf bei“ (2), „auf zu“ (2), „aus mit“ (2), „aus wegen“ (2), „bei in bei“ (2), „bei mit“ (2), „durch in“ (2), „durch mit“ (2), „für in“ (2), „gegenüber in“ (2), „in durch“ (2), „in über“ (2), „in von“ (2), „in von zu“ (2), „mit als“ (2), „mit in“ (2), „nach zu“ (2), „über in“ (2), „unter“ (2), „von auf“ (2), „von aus“ (2), „von bei“ (2), „von über bis“ (2), „zu aus“ (2), „zu für“ (2), „ab“ (1), „ab in“ (1), „als aus“ (1), „als bei“ (1), „als gegen“ (1), „als mit“ (1), „als ohne“ (1), „als wegen“ (1), „an aus“ (1), „an innerhalb“ (1), „an jenseits“ (1), „an über“ (1), „ans in“ (1), „auf als“ (1), „auf als auf“ (1), „auf an in“ (1), „auf an unter“ (1), „auf bei in“ (1), „auf über“ (1), „auf zwischen“ (1), „aufgrund wegen“ (1), „aus als“ (1), „außerhalb zu“ (1), „bei als“ (1), „bei an“ (1), „bei an bei“ (1), „bei auf“ (1), „bei durch“ (1), „bei nach“ (1), „bei unter“ (1), „dank trotz“ (1), „durch auf“ (1), „durch aus“ (1), „durch bei“ (1), „durch für“ (1), „durch nach“ (1), „für aus“ (1), „gegen in“ (1), „gegen über“ (1), „in an in“ (1), „in angesichts“ (1), „in aus“ (1), „in bei als“ (1), „in bei für“ (1), „in bei in“ (1), „in darüber“ (1), „in für gegen“ (1), „in für zu für in“ (1), „in gen“ (1), „in hinichtlich“ (1), „in innerhalb“ (1), „in mit in“ (1), „in übers“ (1), „in um in“ (1), „in unter in“ (1), „in voller“ (1), „in vor“ (1),

„in zwischen“ (1), „innerhalb in“ (1), „innerhalb mit“ (1), „innerhalb vor“ (1), „mit auf“ (1), „mit aufgrund“ (1), „mit bei“ (1), „mit ex ab“ (1), „mit ohne“ (1), „mit per“ (1), „mit über“ (1), „mitsant mit“ (1), „nach als“ (1), „nach an“ (1), „nach vor“ (1), „nahe auf“ (1), „ohne innerhalb“ (1), „ohne zu“ (1), „per“ (1), „per auf“ (1), „über an“ (1), „über auf“ (1), „über bei“ (1), „über durch“ (1), „über nach“ (1), „über um“ (1), „über von“ (1), „um auf“ (1), „um über“ (1), „unter wegen“ (1), „unterhalb auf“ (1), „via mit“ (1), „voller mit“ (1), „von an entlang“ (1), „von bis von bis“ (1), „von in an zu“ (1), „von nach“ (1), „von nach in nach“ (1), „von über“ (1), „vor angesichts“ (1), „vor aus“ (1), „vor nach“ (1), „während an“ (1), „während nach“ (1), „zu als“ (1), „zu ihretwegen“ (1), „zu über“ (1), „zugunsten zu“ (1), „zwischen an in“ (1), „zwischen in“ (1), „zwischen mit“ (1)

2.5.4.2 Funktion der CPP

In der Tabellenzusammenstellung 2.61 auf der nächsten Seite werden die syntaktischen Funktionen von CPP mit PP in NEGRA verglichen. Auffällig sind drei Dinge: Pseudo-Genitiv (PG) ist bei CPP deutlich seltener. Den Gründen dafür kann hier nicht weiter nachgegangen werden. Zweitens sind CPP etwas häufiger Top-Kategorien, d.h. gerne in Titeln und Überschriften anzutreffen. Drittens gibt es mehr CPP in NK-Funktion, was eine Kombination darstellt, welche auf den ersten Blick sehr unsinnig erscheint.

Das Problem wird in Beispiel (133) deutlich: In normalen PP erscheinen Partikeln initial, aber innerhalb der PP:

$$\text{Normal:} \quad [_{PP}[_{ADV-MO} \ldots] \ldots]$$

Auch für CPP gilt die Regel, dass nur Konjunktoren CJ und Konjunkte CD als Tochterkonstituenten erscheinen dürfen. Wenn nun eine CPP als Ganzes durch eine Partikel modifiziert wird, darf diese nicht initial innerhalb der CPP erscheinen:

$$\text{Verboten:} \quad [_{CPP}[_{ADV-MO} \ldots] \ldots]$$

Als Ausweg werden deshalb PP konstruiert, welche aus einer initialen Partikel und einer CPP in NK-Funktion besteht:

$$\text{Ausweg:} \quad [_{PP}[_{ADV-MO}[_{CPP-NK} \ldots]]]$$

.

(133) Arbeitsbeschaffungsprogramme, in denen die notleidende Landbevölkerung [$_{PP}$ [$_{ADV-MO}$ etwa] [$_{CPP-NK}$ beim Straßenbau oder [$_{PP-CJ}$ mit der Anlage von Teichbecken]]] etwas Geld verdienen kann, [...] [N_{4106}]

CPP				PP			
in %	Anzahl	Funktion	kum.	in %	Anzahl	Funktion	kum.
61.6	294	MO	62	61.7	7871	MO	62
22.6	108	MNR	84	26.9	3430	MNR	89
4.0	19	NK	88	4.2	539	PG	93
3.4	16	—	92	3.7	477	CJ	96
1.9	9	CJ	94	1.4	174	SBP	98
1.9	9	APP	95	0.7	91	PD	99
1.7	8	SBP	97	0.7	91	—	99
1.7	8	PD	99	0.3	41	CC	100
0.8	4	PG	100	0.2	24	APP	100
				0.0	4	RE	100
				0.0	4	NK	100

Tabelle 2.61: Verteilung der Funktionen von CPP und PP in NEGRA. Top-Phrasen sind mit '—' ausgezeichnet. Gezeigt werden Fälle mit mindestens 4 Vorkommen.

2.5.5 CAP

Wie in Tabelle 2.1 auf Seite 13 ersichtlich, sind CAP zwar die dritthäufigst vorkommende koordinierte Kategorie, aber in NEGRA enthalten weniger als 1/5 davon mindestens ein phrasales Konjunkt: W (733, 81.4%), P (167, 18.6%).

2.5.5.1 Der Bau der CAP

Die Tabelle 2.62 auf der nächsten Seite zeigt die verschiedenen in NEGRA annotierten Tochterkonstituenten, welche sich auf 49 Types verteilen. Die hohe Anzahl Types wird insbesondere durch die Terminal-Konstituenten verursacht, welche als attributive (ADJA), prädikativ oder adverbiale (ADJD) Adjektive getaggt sind. Auf Grund der Auftrennung der Ziffernschreibweise von grossen Zahlen in Dreiergruppen, welche in NEGRA als einzelne Token behandelt werden, erscheinen die numerischen Mehrwortlexeme (NM) wie in (134) relativ häufig.

(134) Bislang sind auch die dafür nötigen [$_{CAP}$ [$_{NM}$ [$_{CARD-NMC}$ 60] [$_{CARD-NMC}$ 000] [$_{APPR-CD}$ bis] [$_{NM}$ [$_{CARD-NMC}$ 70] [$_{CARD-NMC}$000]]] Mark nicht im Haushalt bereitgestellt [...] [N_{2325}]

In der weitere Diskussion werden die Konstituenten des Typs NM bei Bedarf auf CARD abgebildet, da es sich eher um ein Tokenisierungsproblem handelt.

In der Tabelle 2.63 auf der nächsten Seite mit der kompakten Zusammenstellung von Folgen von Nicht-Terminal und Terminal-Konjunkten zeigt sich, dass die reinen Nicht-Terminalphrasen mit 42% deutlich weniger dominant sind, als dies etwa mit 72% bei den CNP der Fall (siehe Tabelle 2.45 auf Seite 103) ist. Die komplexen Konjunkte befinden sich ebenfalls dominant rechtsseitig.

in %	Anzahl	Tochterkonstituenten	kumulativ
23.4	39	AP KON AP	23
13.8	23	ADJA KON AP	37
10.8	18	ADJD KON AP	48
6.6	11	AP AP	55
4.8	8	NM APPR NM	59
4.2	7	AP KON ADJD	64
3.0	5	ADJD AP	67
2.4	4	AP KON ADJA	69
1.8	3	ADJA AP KON AP	71
1.8	3	ADJD ADJD KON AP	73
1.2	2	ADJA AP	74
1.2	2	AP ADJA KON ADJA	75
1.2	2	AP ADJD	76
1.2	2	AP ADJD KON ADJD	77
1.2	2	AP AP KON AP	79
1.2	2	KON ADJD KON AP	80
1.2	2	NM KON NM	81

Tabelle 2.62: Die Verteilung der lexikalischen Tochterkonstituenten von total 167 CAP in NEGRA. Von den total 49 Typen sind nur diejenigen mit mindestens 2 Vorkommen aufgeführt. Das Type-Token-Verhältnis beträgt 1:3.4 .

in %	Anzahl	Konjunktfolgen	kumulativ
41.9	70	N	42
41.9	70	T N	84
13.8	23	N T	98
1.8	3	T N T	99
0.6	1	T N T N	100

Tabelle 2.63: Die Verteilung der Abfolge von Terminal- und Nicht-Terminalkonstituenten in den Tochterkonjunkten der CAP in NEGRA.
Legende: T = mindestens eine Terminalkonstituente, N = mindestens eine Nicht-Terminalkonstituente

CAP				AP			
in %	Anzahl	Funktion	kum.	in %	Anzahl	Funktion	kum.
45.5	76	NK	46	51.2	1497	NK	51
21.6	36	PD	67	29.5	862	PD	81
21.6	36	MO	89	10.2	299	MO	91
4.8	8	HD	94	4.1	120	CJ	95
1.8	3	MNR	95	1.9	55	MNR	97
1.8	3	—	97	0.9	27	CC	98
				0.7	20	OA	98
				0.5	16	—	99
				0.4	12	SB	99
				0.4	12	APP	100
				0.1	4	OC	100

Tabelle 2.64: Verteilung der Funktionen von CAP und AP in NEGRA. Top-Phrasen sind mit '—' ausgezeichnet. Gezeigt werden Fälle mit mindestens 4 Vorkommen.

2.5.5.2 Funktion der CAP

Gemäss Tabelle 2.64 ist die häufigste Funktion der CAP (und auch der AP) als NK in NP oder PP mit attributiven Adjektiven als Kern. Die Auflistung (135) zeigt alle Kerne[52] von phrasalen Konjunkten von CAP in kondensierter Form, d.h. alle Folgen identischer Wortartentags werden nur durch einen Repräsentanten angezeigt.

(135) Kondensierte Kernfolgen der CAP-NK in NEGRA:
„ADJA“ (52, 68.4%), „CARD“ (17, 22.4%), „ADJD“ (3, 3.9%), „ADJA VZ“ (1, 1.3%), „CAP“ (1, 1.3%), „CAP PIAT“ (1, 1.3%), „NP CARD“ (1, 1.3%)

Etwas auffällig sind die Fälle, wo als Kern ADJD-Folgen angegeben sind. Dies wird bei „als“-Konstruktionen wie in Beispiel (136)[53] gemacht.

(136) Die bisherige, sich hauptsächlich auf das Strafrecht stützende Politik wird von der Kommission [$_{PP}$ als " [$_{CAP}$ [$_{ADJD-CJ}$ inhuman] , [$_{ADJD-CJ}$ wirkungslos] , [$_{ADJD-CJ}$ kontraproduktiv] und [$_{ADJD-CJ}$ schädlich für den Rechtsstaat]] " bezeichnet. [N_{15739}]

Die zweithäufigste Verwendung sind Prädikative (PD), welche fast ausschliesslich ADJD-Kerne aufweisen.

(137) Kondensierte Kernfolgen der CAP-PD in NEGRA:
„ADJD“ (29, 80.6%), „ADJD CAP“ (2, 5.6%), „VVPP ADJD“ (2, 5.6%),

[52] Rekursiv eingebettete CAP werden als Kerne und NM als CARD behandelt.

[53] Dieser Satz ist zugleich ein Beispiel, wo eine koordinative Adjektivreihung nicht als CAP annotiert ist.

„ADJD ADJA“ (1, 2.8%), „ADJD VVPP“ (1, 2.8%), „AP ADJD“ (1, 2.8%)

Eine Standard-Konstruktion findet sich in Beispiel (138a), während der Ausreisser in (138b) eine elliptische rechtsextraponierte CAP diskontinuierlich ans Vorfeld-Konjunkt eingliedert. Da das modifizierende „auch“ Skopus über CAP hat, wird um diese herum noch eine AP gebildet, welche eine CAP als Kopf hat. Solche Fälle stehen hinter den CAP, welche in der Tabelle 2.64 auf der vorherigen Seite mit HD als Funktion auftauchen.

(138) a. [$_{CAP}$ [$_{ADJD}$ Maßgeblich] und [$_{AP}$ in jedem Falle "wasserdicht"]] sind die offiziellen Zahlen des RP. [N$_{4172}$]

b. [$_{CAP}$ [$_{ADJD}$ "Ordentlicher "] ⌊ sei die Schweizer Straße geworden, sagt sie, ⌋ [$_{CD}$ aber] [$_{AP}$ [$_{MO}$ auch] [$_{CAP-HD}$ "fremder" – und "kühler"]]] . [N$_{2023}$]

Die dritthäufigste Verwendung sind Adverbiale (MO), welche fast ausschliesslich ADJD-Kerne haben wie in (140).

(139) Kondensierte Kernfolgen der CAP-MO in NEGRA:
„ADJD“ (31, 86.1%), „ADJD CAP“ (1, 2.8%), „ADJD VVPP“ (1, 2.8%), „CARD“ (1, 2.8%), „PIS ADV“ (1, 2.8%), „VVPP“ (1, 2.8%)

Da bei departizipialen Adjektiven wegen der Formgleichheit mit Partizipien gerne Tagging-Fehler auftreten, sind solche Annotationsfehler verständlich.

(140) [. . .] beim Gehen hielt er die Arme [$_{CAP}$ [$_{AP}$ nicht gerade] , [$_{KON}$ sondern] [$_{AP}$ immer leicht angewinkelt]]. [N$_{13985}$]

Adjektivphrasen in MNR-Funktion sehen problematisch aus. Die 3 CAP bestätigen dies: In (141) werden die Präpositionen „nördlich“ und „südlich“ mit ihren homonymen Adjektiven verwechselt trotz ihrer Kasusrektion.

(141) Das Areal rund um den ehemaligen Schrottplatz [$_{CAP-MNR}$ nördlich des Höllweges und westlich des Park-and-ride-Platzes] ist erheblich belastet [. . .] [N$_{19114}$]

2.5.6 CVP

Wie in Tabelle 2.1 auf Seite 13 ersichtlich, sind CVP die vierthäufigst vorkommende koordinierte Kategorie. In NEGRA enthalten fast alle mindestens ein phrasales Konjunkt: P (532, 97.3%), W (15, 2.7%) .

2.5.6.1 Der Bau der CVP

In der Tabelle 2.66 auf der nächsten Seite mit der kompakten Zusammenstellung von Folgen von Nicht-Terminal und Terminal-Konjunkten zeigt sich, dass die Dominanz der reinen Nicht-Terminalphrasen mit über 83% sehr hoch ausfällt. Die komplexen Konjunkte oft linksseitig.

in %	Anzahl	Tochterkonstituenten	kumulativ
63.0	335	VP KON VP	63
9.2	49	VP VP	72
7.0	37	VP KON VVPP	79
3.4	18	VP KON VVINF	83
2.8	15	VP VP KON VP	85
2.3	12	VP VP VP	88
1.5	8	VVPP KON VP	89
1.1	6	VP KON VZ	90
0.8	4	VP KON CVP	91
0.8	4	VP KON VP KON VP	92
0.8	4	VP KON VVIZU	93
0.6	3	VP VVPP KON VP	93
0.4	2	CVP KON VP	94
0.4	2	KON VP KON VP	94
0.4	2	VP KON KON VP	94
0.4	2	VP KON VP KON VVPP	95
0.4	2	VP VP VP KON VP	95
0.4	2	VVINF KON VP	96
0.4	2	VVPP VVPP KON VP	96

Tabelle 2.65: Die Verteilung der lexikalischen Tochterkonstituenten von total 532 CVP in NEGRA. Von den total 42 Typen sind nur diejenigen mit mindestens 1 Vorkommen aufgeführt. Das Type-Token-Verhältnis beträgt 1:12.7 .

in %	Anzahl	Konjunktfolgen	kumulativ
83.5	444	N	84
12.0	64	N T	96
3.0	16	T N	98
1.3	7	N T N	100
0.2	1	T N T N T N	100

Tabelle 2.66: Die Verteilung der Abfolge von Terminal- und Nicht-Terminalkonstituenten in den Tochterkonjunkten der CVP in NEGRA.
Legende: T = mindestens eine Terminalkonstituente, N = mindestens eine Nicht-Terminalkonstituente

Die in NEGRA annotierten Tochterkonstituenten von CVP-Phrasen sind zu über 60% 2-teilige syndetische Koordinationen von VP-Töchtern. Daneben existieren insbesondere wegen der nicht-finiten Einwort-Konjunkte (Infinitive (VVINF), Partizip Perfekt (VVPP), „zu"-Infinitiv (VZ, VVIZU)) verschiedene Varianten.

Welche nicht-finiten Verbalformen werden nun in den Tochterkonjunkten miteinander kombiniert. Um diese Frage zu beantworten, werden die Köpfe der Konjunkte extrahiert und in kondensierter Form dargestellt, d.h. alle Folgen von identischen Wortarten werden nur durch einen Repräsentanten angezeigt. Dabei werden zusätzlich alle morphologischen „zu"-Infinitive wie beim Wort „aufzuhören" (VVIZU) als VZ aufgefasst. Zudem macht die Unterscheidung von potentiellen Hilfsverben (VA...) und Vollverben (VV...) in diesem Kontext keinen Sinn. Hilfs- und Modalverben werden ebenfalls als wie Vollverben behandelt.

In der Auflistung (142) sind alle kondensierten Abfolgen mit mindestens 3 Vorkommen aufgeführt. Folgen von Konjunkt-Kernen aus Partizip Perfekt sind mit über 1/3 vertreten, etwas weniger als 1/3 betragen Infinitive und knapp 1/4 die „zu„-Infinitive.

(142) Kondensierte Kernfolgen der CVP in NEGRA:
„VVPP" (206, 38.7%), „VVINF" (167, 31.4%), „VZ" (123, 23.1%), „VVPP VP" (6, 1.1%), „VP VVPP" (5, 0.9%), „VP VVINF" (3, 0.6%)

Die Tabellenzusammenstellung 2.67 auf der nächsten Seite zeigt, dass in NEGRA und TIGER die Kernfolgen mit der syntaktischen Funktion der CVP zusammenhängen.

2.5.6.2 Funktion der CVP

In der Tabellenzusammenstellung 2.68 auf Seite 130 sind die häufigsten Funktionen von CVP und VP im Vergleich zu sehen. Dominant sind die sogenannten Objektsätze (OC), mit denen nicht-finite Prädikatsteile mit ihren abhängigen Elementen eingebettet werden wie in (143) auf dieser Seite.

(143) [$_{S}$ [$_{NP-SB}$ Die Leiden des tapferen, der Gewaltlosigkeit verpflichteten Volkes] [$_{VAFIN-HD}$ werden] [$_{CVP-OC}$[$_{VP-CJ}$ auch von unserer Presse zu wenig beachtet] und [$_{VP-CJ}$von unseren Politikern opportunistischen Zielen geopfert]]] . [N$_{7158}$]

Wiederholte Elemente (RE) bei Platzhalterkonstruktionen wie in (144a) mit unpersönlichem Korrelat-„es" sind sowohl in NEGRA wie TIGER etwas häufiger bei den CVP, da schwerere Elemente generell stärker herausgesetzt werden. Konstruktionen mit „zu„-Infinitiven" wie in (144b) erlauben eine kompaktere Formulierung als „dass"-Nebensätze im Passiv.

(144) a. Wie vielen seiner dem Jugendstil anhängenden Kollegen genügt [$_{NP-SB}$ [$_{PPER-PH}$ es] ⌊ Huber nicht ⌋ , [$_{CVP-RE}$[$_{VP-CJ}$ Räume zu gestalten] und [$_{VP-CJ}$ Mobiliar zu entwerfen]]] . [N$_{8376}$]

NEGRA

Funktion	Anzahl	in %	Kernfolge	Anz.	in %
OC	430	80.8	VVPP	189	35.5
			VVINF	148	27.8
			VZ	66	12.4
			VVPP VP	6	1.1
			VP VVPP	4	0.8
RE	30	5.6	VZ	28	5.3
MO	21	3.9	VZ	13	2.4
			VVPP	6	1.1
CJ	11	2.1	VVPP	5	0.9
			VVINF	3	0.6
—	11	2.1	VVINF	5	0.9
			VZ	4	0.8
SB	8	1.5	VZ	6	1.1
PD	6	1.1	VVPP	3	0.6
NK	4	0.8	VVINF	4	0.8

TIGER

Funktion	Anzahl	in %	Kernfolge	Anz.	in %
OC	1048	84.7	VVPP	416	33.6
			VVINF	342	27.6
			VZ	176	14.2
			VP VVINF	25	2.0
			VVPP VP	17	1.4
			VP VVPP	16	1.3
			VVPP VVINF	11	0.9
			VVINF VP	10	0.8
			VP VZ	5	0.4
			CVP VVINF	5	0.4
RE	63	5.1	VZ	60	4.9
MO	43	3.5	VZ	33	2.7
			VVPP	8	0.6
CJ	19	1.5	VVINF	6	0.5
			VZ	5	0.4
			VVPP	5	0.4
PAR	14	1.1	VVPP	7	0.6
PD	12	1.0	VVPP	9	0.7
SB	11	0.9	VZ	9	0.7

Tabelle 2.67: Verhältnis der Funktionen zu den Kernfolgen der CVP in NEGRA und TIGER. Die Spalte „Kernfolge“ ist bei der Diskussion von (142) auf der vorherigen Seite erklärt. Gezeigt sind Zeilen mit mindestens 3 (NEGRA) bzw. 4 (TIGER) Vorkommen.

NEGRA

CVP

in %	Anzahl	Funktion	kum.
80.8	430	OC	81
5.6	30	RE	86
3.9	21	MO	90
2.1	11	CJ	92
2.1	11	—	94
1.5	8	SB	96
1.1	6	PD	97
0.8	4	NK	98
0.8	4	HD	99
0.8	4	APP	100

VP

in %	Anzahl	Funktion	kum.
79.2	9029	OC	79
7.9	897	CJ	87
4.6	526	PD	92
3.4	391	MO	95
2.5	284	RE	98
1.1	131	MNR	99
0.5	53	—	99
0.5	52	SB	100
0.2	27	CC	100
0.1	9	APP	100
0.0	4	NK	100

TIGER

CVP

in %	Anzahl	Funktion	kum.
84.7	1048	OC	85
5.1	63	RE	90
3.5	43	MO	93
1.5	19	CJ	95
1.1	14	PAR	96
1.0	12	PD	97
0.9	11	SB	98
0.6	8	MNR	98
0.6	8	—	99
0.4	5	HD	99

VP

in %	Anzahl	Funktion	kum.
80.8	20550	OC	81
8.7	2223	CJ	90
3.2	819	MO	93
2.9	746	PD	96
2.4	616	RE	98
0.5	139	PAR	98
0.4	104	SB	99
0.3	82	—	99
0.3	77	MNR	100
0.2	49	CC	100
0.1	17	OP	100
0.0	4	PNC	100
0.0	4	NK	100

Tabelle 2.68: Verteilung der Funktionen von CVP und VP in NEGRA und TIGER. Top-Phrasen sind mit '—' ausgezeichnet. Gezeigt werden Fälle mit mindestens 4 Vorkommen.

b. Dabei hatte sich ausgerechnet der zuständige Ortsbeirat vehement [$_{PP}$ [$_{PROAV-PH}$ dafür] ⌊ eingesetzt ⌋ , [$_{CVP-RE}$[$_{VP-CJ}$ die Halle nicht sofort abzureißen] , sondern [$_{VP-CJ}$ für Ausstellungen oder andere kulturelle Ereignisse zu nutzen]]] . [N_{9235}]

Warum prädikative Verwendung (PD) wie in (145) bei CVP in NEGRA und TIGER seltener sind als bei VP, muss hier offen bleiben.

(145) Keine einzige Konstellation der Figuren ist [$_{CVP-PD}$ [$_{VP}$ zur Sinnfälligkeit entwickelt] oder [$_{VP}$ auch nur zur Anschaulichkeit gebracht]] . [T_{5501}]

2.5.6.3 Ellipsen in CVP

Elliptische Auslassungen sind auch in den Konjunkten von CVP zu finden. Da im TIGER-Korpus dieses Auslassungen durch sekundäre Kanten ergänzt werden, können diese Phänomene nur dort automatisch erhoben und quantifiziert werden.

Wie viele Koordinationen sind überhaupt elliptisch? Die Auflistung (146) zeigt, bei wie vielen Koordinationen wie viele Konjunkte durch sekundäre Kanten ergänzt werden. 2/3 der CVP gelten gemäss Annotation nicht als elliptisch, bei 1/3 gibt es mindestens 1 Konjunkt, das ergänzt wird mit Material aus andern Konjunkten.

(146) Verteilung der CVP in TIGER, welche *n* Konjunkte mit elliptischem Material haben:
0 (837, 67.7%), 1 (358, 28.9%), 2 (38, 3.1%), 3 (4, 0.3%)

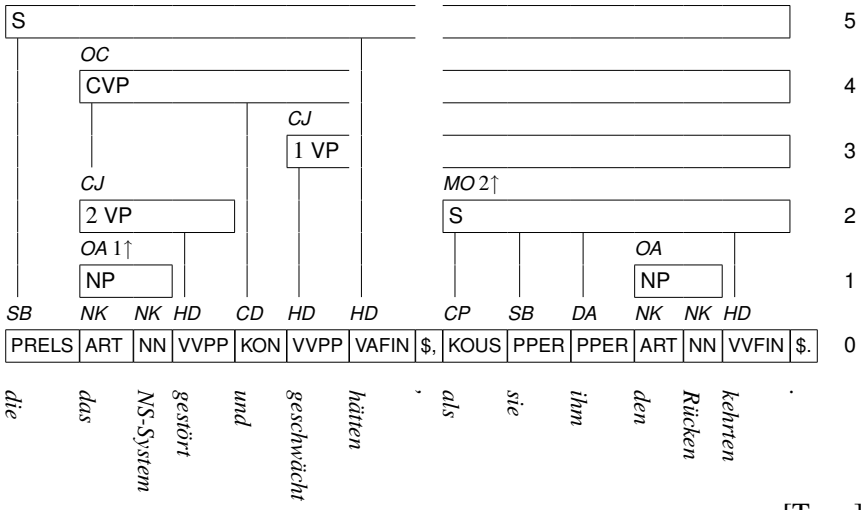
In welchen Verhältnis stehen nun bei den elliptischen Koordinationen die Anzahl der nicht-elliptischen Konjunkte zu den elliptischen? Die Auflistung (147) ergibt als dominante Form je 1 vollständiges und 1 elliptisches Konjunkt.

(147) Verteilung der Anzahl Konjunkte in TIGER dargestellt als Verhältnis nicht-elliptisch:elliptisch:
1:1 (345, 86.2%), 1:2 (20, 5.0%), 0:2 (16, 4.0%), 2:1 (10, 2.5%), 3:1 (3, 0.8%), 1:3 (3, 0.8%), 2:2 (2, 0.5%), 2:3 (1, 0.2%)

Auffällig sind die Verhältnisse 0:2, welche besagen, dass beide Konjunkte elliptisch sind. Darunter fallen Sätze wie (148a)[54], wo eine enge Kontakt-Koordination nicht auf der Ebene der Partizipien koordiniert wird. Aber auch Verb-Ellipsen wie in (148b), wo einerseits eine Verbalkopf-Ellipse und eine Akkusativobjekt-Ellipse vorliegt.

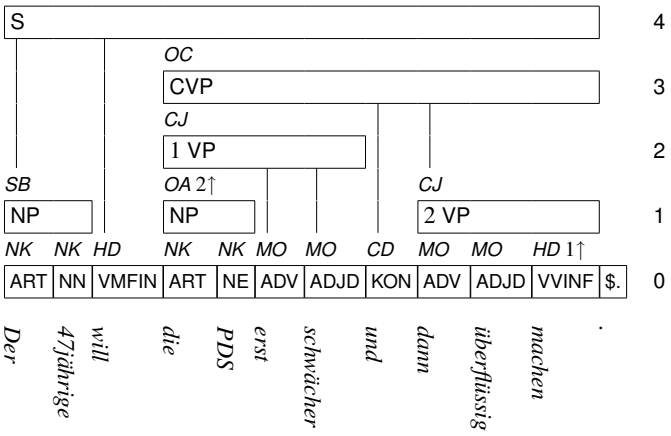
(148) a. Die Berliner Ärztin Gertrud Gumlich forderte auf der EKD-Synode Gerechtigkeit für diese Gruppe Namenloser ,

[54] Sekundäre Kanten werden in Kastendiagrammen über Indizierung dargestellt: $n\uparrow$ bei einer funktionalen Bestimmung markiert den Beginn einer Kante, welche bei der mit *n* indizierten Konstituente endet.



[T_{8293}]

b.



[T_{17557}]

Welche syntaktischen Funktionen sind nun besonders betroffen? Die Auflistung (149) zeigt, dass etwas mehr als 1/2 der total 595 elliptischen Konstituenten als Adverbiale fungieren (MO), 1/5 als Akkusativobjekte (OA) und knapp 15% der elliptischen Konstituenten betreffen den Verbal-Kopf.

(149) Verteilung der Funktionen der einzelnen elliptischen Konstituenten: MO (317, 53.3%), OA (128, 21.5%), HD (85, 14.3%), CP (27, 4.5%), SBP (18, 3.0%), NG (7, 1.2%), DA (6, 1.0%), PD (3, 0.5%), OP (3, 0.5%), CM (1, 0.2%)

Wie Beispiel (150) zeigt, werden nebensatzeinleitende Konjunktionen (KOUI) durch sekundäre Kanten ergänzt, obwohl eine Einbettung auf der primären Strukturebene als $[_{KOUI}\ [_{CVP}\ \ldots]]$ leicht möglich wäre.

(150) Er habe schon seit längerer Zeit befürchtet, " daß irgendein Rattenfänger aus der rechtsradikalen Ecke sich dieses Themas bemächtigt, $[_{CVP}\ [_{VP}\ [_{KOUI-CP\ 1\uparrow}$ um] Angst zu schüren] und $[_{1\ VP}$ Stimmen zu fangen]] ", [...] [T_{4518}]

Zuletzt soll noch betrachtet werden, welche Funktionskombinationen an die total 446 elliptischen Konjunkt gehängt werden, da bei einigen Konjunkten mehr als 1 Konstituente fehlt. Solche mehrfach elliptischen Konjunkte sind in Auflistung (151) zusammengestellt. Ein Eintrag wie „MO<MO<OA“ (6, 1.3%) besagt: Es gibt 6 Fälle, wo ein Konjunkt um 2 Adverbiale (MO) und ein Akkusativobjekt (OA) ergänzt werden.

(151) Verteilung der ergänzten Funktion(skombination)en pro Konjunkt: „MO“ (131, 29.4%), „OA“ (82, 18.4%), „HD“ (70, 15.7%), „MO<MO“ (41, 9.2%), „MO<OA“ (27, 6.1%), „CP“ (21, 4.7%), „SBP“ (13, 2.9%), „MO<MO<MO“ (11, 2.5%), „MO<MO<OA“ (6, 1.3%), „HD<OA“ (6, 1.3%), „MO<NG“ (4, 0.9%), „HD<MO“ (4, 0.9%), „CP<OA“ (4, 0.9%), „MO<SBP“ (3, 0.7%), „DA“ (3, 0.7%), „NG“ (2, 0.4%), „MO<MO<SBP“ (2, 0.4%), „MO<MO<MO<MO“ (2, 0.4%), „PD“ (1, 0.2%), „OP“ (1, 0.2%), „OA<OP“ (1, 0.2%), „MO<OP“ (1, 0.2%), „MO<MO<MO<OA“ (1, 0.2%), „HD<PD“ (1, 0.2%), „HD<MO<PD“ (1, 0.2%), „HD<MO<OA“ (1, 0.2%), „HD<MO<MO“ (1, 0.2%), „DA<MO“ (1, 0.2%), „DA<HD“ (1, 0.2%), „CP<NG“ (1, 0.2%), „CP<DA“ (1, 0.2%), „CM“ (1, 0.2%)

2.5.7 CO

Diese Kategorie dient als Sammelbecken für alle Koordinationen, deren Tochterkonjunkte auch in elliptischer Lesart nicht mit derselben Phrasenkategorie gebaut werden können, und so tauchen darin auch die unterschiedlichsten Kombinationen auf. Die echte Zahl an CO dürfte etwas höher sein, da diese Kategorie nicht immer gleich konsistent und zwingend angewendet wurde. Insbesondere bei den CS sind einige nicht-verbale Absolutkonstruktionen als S annotiert. CO ist vorwiegend eine Phrasenkategorie, wie die folgende Verteilung zeigt: P (158, 95.2%), W (8, 4.8%)

Die in NEGRA annotierten Tochterkonstituenten von CO-Phrasen sind extrem vielfältig. Die Tabelle 2.69 auf der nächsten Seite zeigt, dass mehr als 1/3 aller Phrasenkomponenten nur 1 Vorkommen haben. Entsprechend tief ist das Type-Token-Verhältnis von 1:1.9 .

Die häufigste Form von CO verknüpft ein ADJD syndetisch mit einer PP in Modifikatorfunktion, welche dieselbe semantische Adverbialfunktion ausüben. Dies lässt sich durch Fragetests operationalisieren. In Beispiel (152a) oder (152b) ist es mit dem Fragewort „Wie?“ die Art und Weise. Etwas weniger klar vielleicht in (152c), wo das „wie“ mit einem metaphorischen „wohin“ verknüpft wird. In (152d) mit einem Lokaladverbiale.

(152) a. [$_{CO-MO}$ Ungeduldig und mit Spannung] erwarten - wie sicherlich so mancher Kulturtourist - aber auch einige Künstler den Eröffnungstag: [...] [N$_{15921}$]

b. Der Kreistag beschloß [$_{CO-MO}$ einstimmig und ohne Diskussion] , [...] [N$_{16753}$]

in %	Anzahl	Tochterkonstituenten	kumulativ
6.3	10	ADJD KON PP	6
5.7	9	PP KON NP	12
5.1	8	NP KON S	17
4.4	7	NP KON PP	22
4.4	7	S NP	26
3.2	5	ADJD KON VP	29
3.2	5	AP KON PP	32
3.2	5	PP KON AP	36
2.5	4	PP KON VP	38
1.9	3	AP KON S	40
1.9	3	NN KON AP	42
1.9	3	NP KON AP	44
1.9	3	PP KON S	46
1.9	3	VP KON PP	48
1.3	2	ADJD KON NP	49
1.3	2	ADV KON PP	50
1.3	2	AP KON VP	51
1.3	2	AVP KON PP	53
1.3	2	CNP KON PP	54
1.3	2	CNP S S	55
1.3	2	NP KON ADJD	57
1.3	2	NP S	58
1.3	2	PP KON ADJD	59
1.3	2	PP KON ADV	60
1.3	2	S KON VP	62
1.3	2	VP KON AP	63
1.3	2	VP KON S	64

Tabelle 2.69: Die Verteilung der lexikalischen Tochterkonstituenten von total 158 CO in NEGRA. Von den total 84 Typen sind nur diejenigen mit mindestens 2 Vorkommen aufgeführt. Das Type-Token-Verhältnis beträgt 1:1.9 .

in %	Anzahl	Konjunktfolgen	kumulativ
64.6	102	N	65
25.3	40	T N	90
8.9	14	N T	99
1.3	2	N T N	100

Tabelle 2.70: Die Verteilung der Abfolge von Terminal- und Nicht-Terminalkonstituenten in den Tochterkonjunkten der CO in NEGRA: T = mindestens eine Terminalkonstituente, N = mindestens eine Nicht-Terminalkonstituente

c. [...] , steigen die Preise [$_{CO-MO}$ unaufhaltsam und ins Unerschwingliche] . [N$_{4087}$]

d. Der restliche Platz sei für Radwege, Trottoirs und Pflanzungen [$_{CO-MNR}$ [$_{ADJD}$ seitlich] und [$_{PP}$ in der Mitte zwischen den vier Fahrbahnen]] vorgesehen. [N$_{10212}$]

Adverbiale in gleicher Funktion können auch bei „PP KON S“ vorliegen wie in (153a), wo gemäss (Dudenredaktion 2005, §1193) eine „konditionale“ Angabe vorliegt, welche mit „In welchem Fall?“ erfragt werden kann. Bei (153b) wird dieselbe semantische Funktion durch Pseudo-Genitiv mit „von“ und normalem postnominalen Genitiv ausgedrückt.

(153) a. "Wir wollen einen Raum, wo wir uns auch [$_{CO}$[$_{PP}$ bei Regen] ⌊ treffen können ⌋ , und [$_{S}$ wenn es kalt ist] ", verlangte er. [N$_{4172}$]

b. Die Verpflichtung [$_{CO-GR}$ [$_{PP}$ von Florian Weichert und Michael Spies] , sowie [$_{NP-CJ}$ des Vertragsamateurs Thomas Lässig vom Absteiger Hansa Rostock]] kostet [...] [N$_{15384}$]

Teilweise werden damit auch ungrammatische Konstruktionen sanktioniert, welche durch fehlgeleitete Auslassung entstanden sind.

(154) Die Erschließung [$_{CO-MNR}$ [$_{PP}$ mit Ver- und Entsorgungsleitungen] sowie [$_{NP}$ der Zufahrt]] seien gesichert. [N$_{17197}$]

Funktionen der CO Die Tabelle 2.71 auf der nächsten Seite zeigt, dass die für Adverbiale typischen Funktionen MO und MNR etwas über 1/3 abdecken. Weiter sind CO als Top-Phrasen recht häufig.

2.5.8 CAVP

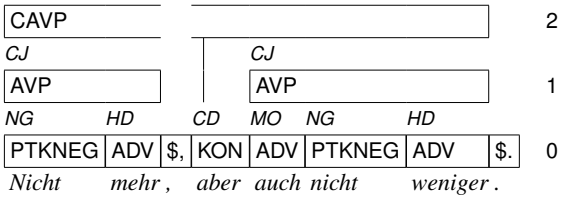
Wie in Tabelle 2.43 auf Seite 101 ersichtlich, sind Wortkoordinationen bei CAVP fast 10 Mal häufiger als die wenigen existierenden 9 Phrasenkoordinationen in NEGRA. Recht prominent ist mit 2 Vorkommen die Wendung „nicht mehr und nicht weniger“: In (155a) als Adverbialphrase in Subjektfunktion annotiert. In (155b) als isoliertes Satzfragment. Die Negationspartikel spielt auch in (155c) oder (155d) mit, sowie in der Floskel „nicht immer, aber immer öfter“ in (155e). Die meisten koordinierten CAVP sind adversative Kombinationen.

(155) a. Was not tut, ist nach Ansicht des Bundesbank-Direktoriumsmitglieds [$_{AVP-SB}$ [$_{CAVP-HD}$ [$_{AVP-CJ}$ [$_{PTKNEG-NG}$ nicht] [$_{ADV-HD}$ mehr] und [$_{PTKNEG-NG}$ nicht] [$_{ADV-HD}$ weniger]]] , [$_{VP-CC}$ als das politische System von seiner Neigung zu übermäßiger Kreditaufnahme zu befreien]]. [N$_{20184}$]

in %	Anzahl	Funktion	kumulativ
29.7	47	MO	30
19.0	30	—	49
13.3	21	PD	62
8.2	13	APP	70
7.0	11	NK	77
7.0	11	MNR	84
3.2	5	OC	87
1.9	3	SB	89
1.9	3	OA	91
1.9	3	GR	93
1.3	2	RS	94
1.3	2	RE	96
1.3	2	PG	97
1.3	2	CJ	98
0.6	1	RC	99
0.6	1	HD	100
0.6	1	CC	100

Tabelle 2.71: Verteilung der Funktionen von CO in NEGRA. Top-Phrasen sind mit '—' ausgezeichnet.

b.



[N_{6433}]

c. Michael Gielens Kompositionen teilen das Schicksal von Adornos Kompositionen, [$_{CAVP}$ nicht oder höchst selten] aufgeführt zu werden. [N_{20184}]

d. Ob es sich tatsächlich [$_{CAVP-MO}$ [$_{ADV-CJ}$ so] [$_{KON}$ und] [$_{AVP-CJ}$ nicht anders]] zugetragen hat, [...] [N_{8725}]

e. [...] , kommt mir – [$_{CAVP-MO}$ [$_{AVP-CJ}$ nicht immer] , [$_{KON}$ aber] [$_{AVP-CJ}$ immer öfter]] – das Grauen. [N_{8486}]

2.6 Mit der Koordination verwandte oder verwechselbare Konstruktionen

2.6.1 Apposition

2.6.1.1 Enge und lockere Apposition nach Duden

Im Grammatik-Duden (Dudenredaktion 2005, §1550) wird die Apposition als Oberbegriff wie folgt gekennzeichnet:

- Funktional: Appositionen sind Attribute (Gliedteil einer Nominalphrase), die von einem Nomen oder einer Nominalphrase abhängen.
- Formal: Sie bestehen aus einer Nominalphrase oder einem Nomen. Mit „wie“ oder „als“ eingeleitete Appositionen gibt es also nicht, ebensowenig wie Appositionen, welche ein Präpositionalphrase bilden.
- Morphologisch: Appositionen stehen entweder im gleichen Kasus wie ihre Bezugsphrase oder das Bezugsnomen, oder sie sind morphologisch unmarkiert (Nominativform). Im Detail ergeben sich allerdings viele Spezialfälle.

Die Apposition wird normalerweise in zwei Hauptgruppen aufgeteilt:

- Lockere Apposition, welche im Schriftsystem durch paariges[55] Komma abgetrennt wird, und manchmal rechtsextraponiert im Nachfeld erscheint. In der mündlichen Sprache wird insbesondere vor der Apposition stimmlich eine Absetzung gemacht. Gemäss (Dudenredaktion 2005, §1552) „erläutert oder identifiziert“ die Apposition die Phrase, zu der sie gehört.
- Enge Apposition, welche nicht durch Komma abgetrennt und nicht rechtsextraponierbar ist. Darunter fallen bei näherer Betrachtung allerdings syntaktisch und semantisch sich stark unterscheidende Konstruktionen

Lockere Appositionen sind ihrem Bezugsnomen immer nachgestellt. Die Verwendung des Kommas kann lockere Appositionen oberflächlich betrachtet ähnlich zu asyndetischen bzw. monosyndetischen Koordinationen machen.

Die Kasuskongruenz zwischen Apposition und Bezugsgrösse ist grammatisch dann verlangt, wenn in der Apposition ein Begleiter den Appositionskern modifiziert. Sie ist üblich, wenn attributive Adjektive dazu kommen. Wenn keine flektierte Kategorie vor dem Appositionskern vorhanden ist, wird die Nominativform verwendet. Insbesondere artikellose Eigenname zeigen keine Genitivendung in diesem Kontext.

Bei Präpositionen, welche zwischen Genitiv und Dativ schwanken, sowie beim Pseudo-Genitiv mit „von“ besteht eine von der normativen Grammatikschreibung nicht akzeptierte Tendenz, die Fälle zu vermischen. Es gibt aber auch eine sprachpflegerisch nicht geschätzte Tendenz, den Dativ als „neutralen Appositionskasus“ zu verwenden.

[55] Am Satzende fällt das schliessende Komma natürlich weg.

in %	Anzahl	Apposition	kumulativ
68.2	1359	NP-APP	68
8.4	168	CNP-APP	77
8.2	164	MPN-APP	85
7.7	154	S-APP	92
2.7	54	PP-APP	95
1.4	27	NM-APP	97
0.8	16	AP-APP	97
0.7	14	CO-APP	98
0.6	11	VP-APP	99
0.5	9	CPP-APP	99
0.4	8	CS-APP	100
0.2	4	CVP-APP	100
0.2	4	CAP-APP	100
0.1	1	CAVP-APP	100

Tabelle 2.72: Verteilung der total 1993 Konstituenten mit APP-Funktion in NEGRA

2.6.1.2 Die lockere Apposition in NEGRA

Das Annotationshandbuch von NEGRA (Brants u. a. 1999, 21) sieht bei den NP und nur bei den NP lockere Appositionen vorliegen, wenn sie durch Kommata oder Klammern vom Bezugsnomen abgetrennt sind. Insbesondere dann, wenn zwischen Apposition und Bezugnomen noch Genitivattribute vorhanden sind. Nichtsdestotrotz gibt es im NEGRA-Korpus doch eine vielfältige Sammlung von Konstituenten, welche als Apposition markiert sind, aber keine nominale Konstituente darstellen. Die Tabelle 2.72 gibt eine Aufschlüsselung zur Häufigkeit. Darin sieht man, dass knapp 85% der annotierten Appositionen nominal sind.

Die Abgrenzungsmittel der Apposition sind für das NEGRA-Korpus in der Tabelle 2.73 auf der nächsten Seite aufgeführt. Das Komma $, sowie $(, d.h. die satzinterne Interpunktion, sind dominant. $. auf der linken Seite steht für Doppelpunkt[56], „—“ zeigt den Satzbeginn bzw. das Satzende an, „TAG“ steht für ein beliebiges Nicht-Interpunktionstag eines lexikalischen Wortes.

Die lockere Apposition ist oben als der Bezugsgrösse nachgestellte Konstruktion beschrieben worden. Auffällig sind deshalb die Appositionen, welche direkt links an der Satzgrenze stehen. Bei den NP gibt es drei annotierte Exemplare, wovon jedes seine eigenen Fragen aufwirft. Die als doppelte lockere Apposition annotierte Konstruktion in (156a) scheint mir eher ungrammatisch. Interpretierbar wird sie am ehesten, wenn man sich die beiden Appositionen als koordinierte Parenthese vorstellt wie in (156b), mit der eine so nicht mögliche Topikalisierung versucht wurde. Wenn die beiden Zeitadverbien nicht den Funktionsbezeichnungen beige-

[56] Eine detaillierte Untersuchung zur Verwendung des Doppelpunkts liegt mit Karhiaho (2003) vor.

in %	Anzahl	Tag links	Tag rechts	kumulativ
40.6	809	$,	$,	41
28.7	571	$(	$(	69
12.8	255	$,	$.	82
5.4	107	$,	$(	88
4.1	82	$.	$.	92
2.7	54	$(	$.	94
1.5	30	$,	TAG	96
0.8	15	$(	$,	97
0.6	11	$.	$(	97
0.5	9	TAG	TAG	98
0.4	8	TAG	$(	98
0.4	7	TAG	$.	98
0.3	6	—	$(	99
0.3	6	$.	—	99
0.3	5	TAG	$,	99
0.3	5	$,	—	100
0.2	3	$(	TAG	100
0.2	3	$(	—	100
0.1	2	—	$.	100
0.1	2	—	$,	100
0.1	2	$.	$,	100
0.1	1	—	TAG	100

Tabelle 2.73: Aufschlüsselung der unmittelbar angrenzenden Tags bei Konstituenten mit APP-Funktion in NEGRA. „—“ steht für Satzbeginn bzw. Satzende (ohne Interpunktion). „TAG“ steht für ein beliebiges lexikalisches Kürzel.

stellt wären, d.h. „Datenschützer und Innenminister Bull meint“, wäre der Satz unproblematisch und „Bull“ eine enge Apposition.

In Beispiel (156c) wird ebenfalls eine Topikalisierung annotiert – die rhetorische Kontrastbildung von „[stellen] für viele Bürger eine Augenweide [dar]“ mit „[stellen] für andere Menschen jedoch ein Problem dar“ geht dabei nicht in die Analyse ein.

Im Beispiel (156d), das absolut akzeptabel ist, liegen mit der abgesetzten Hervorhebung von „124 000 Mark“ am Satzanfang unklare syntaktische Verhältnisse vor, welche stark mit dem demonstrativen Begleiter zusammenhängen – denn man kann ihn nicht durch einen normalen Artikel ersetzen. Auch wenn man mit „124 000 Mark“ das Vorvorfeld (Thim-Mabrey 1988) besetzt sieht und auch die damit gerne verbundene Hörerlenkung, bleibt die Frage bestehen, ob und in welcher Form die Besetzung dieses Felds an Satzglieder anzuhängen ist.

(156) a. [$_{NP}$ [$_{NP-APP}$ Ehemals Datenschützer] , [$_{NP-APP}$ heute Innenminister] Bull] meint: [N$_{4172}$]

b. Bull, ehemals Datenschützer, heute Innenminister, meint:

c. [$_{NP}$ [$_{NP-APP}$ Für viele Bürger eine Augenweide] , ⌊stellen⌋ über den Gartenzaun auf den Bürgersteig ragende Äste und Zweige von Bäumen und Büschen] für andere Menschen jedoch ein Problem dar. [N$_{7302}$]

d. [$_{NP}$[$_{NP-APP}$ 124 000 Mark] - [$_{NP}$ diesen Betrag]] gilt es zu überbieten, bei der Caritas-Sammelwoche, die am heutigen Freitag beginnt. [N$_{17919}$]

Eine manuelle Durchsicht der über 150 Satz-Appositionen (S-APP), welche aber in 83 Fällen keinen expliziten verbalen Kopf haben, zeigt, dass sie durchwegs Parenthesen darstellen, welche gemäss TIGER-Annotationskonventionen mit PAR zu annotieren wären.

PP-Appositionen Die PP-Appositionen vereinigen recht unterschiedliche Funktionen in sich. In Beispiel (157a) liegt eigentlich eher eine Nebenordnung der beiden referenzidentischen „über“-PP vor. Gleichzeitig ist es ein rhetorisches Mittel, das Thema erst am Schluss abgesetzt durch Komma oder oft auch Doppelpunkt im Nachfeld des Satzes nochmals aufzunehmen. Eine weitere Kategorie sind normale postnominale Modifikatoren, welche in Klammern geschrieben sind, um den Hintergrundcharakter oder kommentierenden/illustrierenden Charakter der Information zu betonen wie in (157b). Am häufigsten werden in Klammern Zusatzinformationen geliefert, welche normalerweise mit einem Prädikat erweitert werden müssen, um interpretiert werden zu können, d.h. als elliptische Parenthesen gelten. Im NEGRA-Korpus gibt es insbesondere in den annotierten Veranstaltungshinweisen aus dem Kulturteil oft Zeit bzw. Ortsangaben in Klammern. Von den 54 PP-Appositionen erscheinen 27, d.h. genau die Hälfte, in runden Klammern.

(157) a. Nicht zufälligerweise verfaßt er seine wohl gelungenste Kunststudie noch in den sechziger Jahren über den Spätromantiker an der Schwelle zur Moderne, [$_{PP-APP}$ über Gustav Mahler] . [N$_{1832}$]

b. Ironische , hintersinnig beziehungsreiche Zeichnungen von Simon E. Waßermann ([$_{PP-APP}$ aus dessen " Tagebüchern der 80er Jahre] "), [. . .] [N_{2403}]

2.6.1.3 Die lockere Apposition in TIGER

Bei den Annotationsrichtlinien bezüglich Appositionen ergeben sich deutliche Änderungen von NEGRA zu TIGER, welche insbesondere mehr Parenthesen-Annotation verlangen. Im TIGER-Annotationshandbuch (Albert u. a. 2003, 23) werden zwei Hauptkriterien benannt, welche Appositionen erfüllen müssen:

- Syntaktische Integration: Eine Apposition muss anstelle ihrer Bezugskonstituente stehen können, wobei nur die Beifügung von Artikeln bei artikellosen Appositionen, sowie allfällige morpho-syntaktischen Anpassungen erlaubt sind.

- Koreferenz: Die Apposition und die Bezugsgrösse müssen das Gleiche bezeichnen.

Abgesehen von diesen operationalisierbaren Tests werden Appositionen auf NP sowie in seltenen Fällen auf PP eingeschränkt. Damit wird insbesondere den in schriftlichen Texten in Klammern beigefügten Zusatzinformationen wie in Beispiel (158) der Status der Apposition abgesprochen.

(158) Der Veba-Konzern ([$_{CNP-APP}$ [$_{NP-CJ}$ 59 Milliarden Mark Umsatz] , [$_{NP-CJ}$ rund 117000 Beschäftigte]]) will trotz der Konjunkturschwäche gut sechs Milliarden Mark pro anno investieren. [N_{1039}]

2.6.2 Linksherausstellung

Im Deutschen ist es möglich, satzgliedwertige Strukturen als Thematisierungsausdrücke (Zifonun u. a. 1997, 518) links vor das Vorfeld zu stellen[57]. Die Sprechpause wird graphematisch durch Komma, Gedankenstrich oder auch Doppelpunkt ausgedrückt. Die Abbildung 2.11 auf der nächsten Seite zeigt einen hochkomplexen Syntaxgraphen aus NEGRA mit Linksherausstellung eines Nebensatzes. Solche Strukturen werden in NEGRA auf der primären Stufe syntaktisch integriert, indem das herausgestellte Element mittels Platzhalterfunktion (PH) an sein Korrelat angehängt wird, welches die Funktion RE (wiederholtes Element) trägt. Durch die deiktische Wiederaufnahme der Linksherausstellung im Satz, lässt sich diese Konstruktion semantisch gut identifizieren, für eine flache syntaktische Analyse hingegen stellen sich Abgrenzungsprobleme.

[57] Die Arbeit von Altmann (1981) stellt die grundlegende Arbeit zu diesen Phänomenen dar.

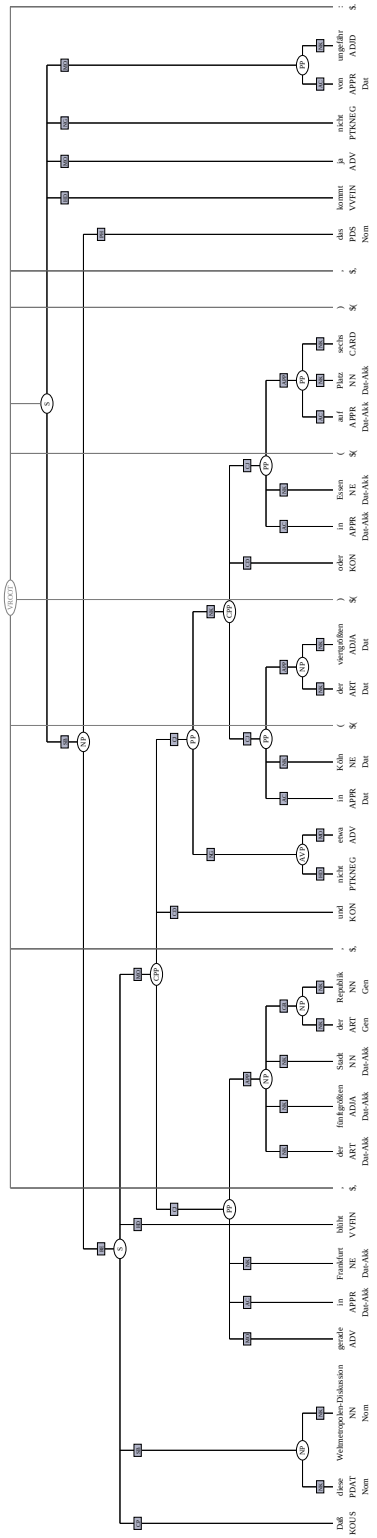


Abbildung 2.11: Linksherausstellung eines Nebensatzes in NEGRA

```
(SBARQ (WHNP-1 Who)
       (SQ (NP-SBJ *T*-1)
           (VP threw
               (NP the ball)))
       ?)

(SBARQ (WHNP-2 What)
       (SQ did
           (NP-SBJ Casey)
           (VP throw
               (NP *T*-2)))
       ?)

(SBARQ (WHNP-3 Who)
       (SQ (NP-SBJ *T*-3)
           (VP will
               (VP throw
                   (NP the ball))))
       ?)
```

Abbildung 2.12: Beispiele für Lücken in der Penn-Treebank nach Bies u. a. (1995): Die Lücken sind mit *T* auf der Terminalebene markiert und dort mit *-N* indiziert. Die Lückenfüller sind Phrasen, welche denselben Index *-N* an ihrer Kategorie aufweisen. Das 1. und 3. Beispiel zeigt die Subjektlücke, welche bei Subjektfragen annotiert wird. Das 2. Beispiel die Objektlücke bei Objektfragen.

2.7 Diskontinuität in koordinierten Strukturen

Das Annotationsmodell von NEGRA erlaubt es, auf einfachste Weise diskontinuierliche Strukturen auszudrücken. Dies im Gegensatz etwa zur *Penn-Treebank*, der wichtigsten Baumbank für das Englische, welche im dazugehörigen umfangreichen Annotationshandbuch (Bies u. a. 1995) detaillierte syntaktische Beschreibungen enthält, wo die unterschiedlichen Spuren anzusetzen sind. Abbildung 2.12 zeigt Beispiele für Spuren, welche bei Fragesätzen entstehen.

In Skut u. a. (1997, 98) wird die Äquivalenz der Strukturbeschreibung mittels Spuren und diskontinuierlichen Konstituenten formuliert als „A context-free constituent backbone can still be recovered from the surface string and argument structure by reattaching 'extracted' structures to a higher node“. In Abbildung 2.13 auf der nächsten Seite ist diese Äquivalenz an einem einfachen Satz illustriert, der zeigt, wie die Strukturumwandlung durch das Werkzeug `negra-tocfg` erfolgt, das Teil des NEGRA-Annotationssoftwarepakets[58] ist.

[58]Erhältlich via `http://www.coli.uni-saarland.de/projects/sfb378/negra-corpus/negra-corpus.html`.

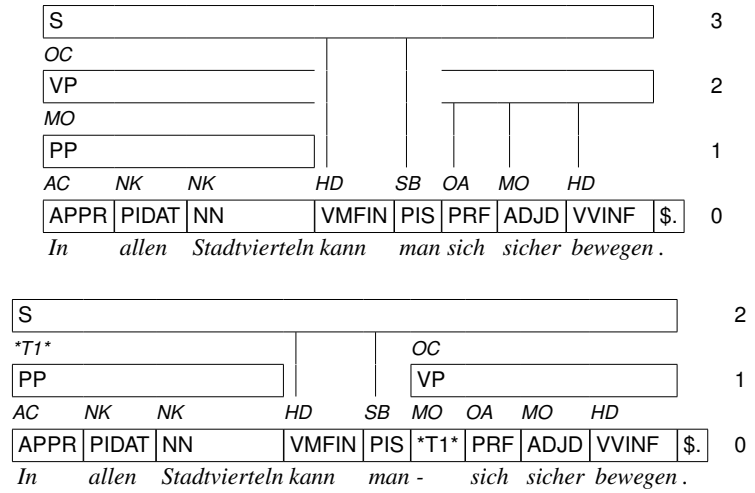


Abbildung 2.13: NEGRA-Baum mit diskontinuierlicher Konstituente (oben) und mit der automatisch daraus erzeugten kontextfreien Repräsentation mit Spuren (unten). Man beachte, dass die Spur in NEGRA auf der Wortarten-Ebene referenziert wird, während die Lückenfüller-Position auf der funktionalen Ebene markiert ist.

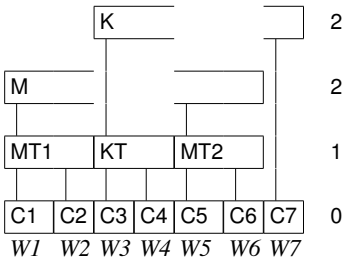
2.7.1 Typen von Diskontinuierlichkeit

Bei der Untersuchung von Diskontinuierlichkeit, wie sie sich in diskontinuierlichen Konstituentenstrukturen manifestieren, sollen folgende Typen unterschieden werden:

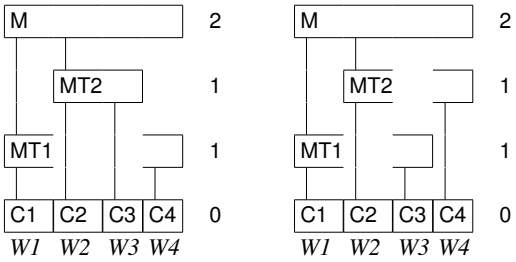
1. Reine Mutter-Diskontinuierlichkeit (RMD): Dies betrifft die Fälle, wo konstituentenfremdes Material nur zwischen vollständigen Tochterkonstituenten eingefügt wird. D.h. nur die Mutterkonstituente ist diskontinuierlich, alle Tochterkonstituenten sind kontinuierlich. Als allgemeine Regel formuliert, wo mögliches diskontinuierliches Material mittels $\lfloor \ldots \rfloor$ markiert ist:

   $$M \rightarrow T_1 \lfloor \ldots \rfloor T_2 \lfloor \ldots \rfloor T_3 \ldots T_n$$

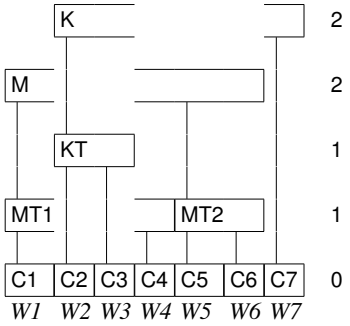
   In Form eines Kastendiagramms ergibt sich für eine einfache RMD-Konfiguration somit folgende Struktur:



2. Reine Tochter-Diskontinuierlichkeit (RTD): Die betrifft die Fälle, wo die Mutter kontinuierlich ist, aber mindestens eine Tochterkonstituente ist von einer andern Tochterkonstituente (oder einem Teil einer andern Tochterkonstituente) unterbrochen. Der Fall, dass eine Nicht-Tochter eine Tochter unterbrechen kann, ist nur möglich, wenn auch die Mutter diskontinuierlich ist. Die beiden möglichen Fälle, d.h. Einschluss einer Tochter oder Verzahnen von mindestens 2 Töchtern ist im Folgenden dargestellt.



3. Mutter-Tochter-Diskontinuierlichkeit (MTD): Dies betrifft die Fälle, wo konstituentenfremdes Material die Mutter und auch mindestens eine Tochterkonstituente unterbricht. Als Kastendiagramm ergibt sich folgende Struktur:



4. Mutter-Diskontinuierlichkeit (MD): Dies betrifft alle Fälle, wo konstituentenfremdes Material die Mutter unterbricht. Es ist die Vereinigungsmenge von RMD und MTD. Diese Kategorie entspricht das Knoten-Prädikat „discontinuous()“ in der TIGER-Search-Abfragesprache (vgl. Abschnitt 6.3.1 auf Seite 276).

Interpunktionszeichen, welche in NEGRA nicht in die Konstituentenstruktur eingebaut werden, sondern als isolierte Terminale stehen bleiben und in der Kastendiagrammdarstellung die Konstituenten visuell ebenfalls unterbrechen, erzeugen keine diskontinuierlichen Konstituenten.

Wie finden sich nun diskontinuierliche Strukturen im NEGRA-Korpus, das im Gegensatz zur Penn-Treebank nur über ein rudimentäres Annotationshandbuch verfügt, das solche Phänomene eingehend beschreibt?

Typ	Anzahl	in %	± kontinuierlich	Anzahl	in %
CNP	5169	52.0	+	5118	51.5
			-	51	0.5
CS	2555	25.7	+	2480	24.9
			-	75	0.8
CAP	900	9.1	+	889	8.9
			-	11	0.1
CVP	547	5.5	+	439	4.4
			-	108	1.1
CPP	477	4.8	+	448	4.5
			-	29	0.3
CO	166	1.7	+	159	1.6
			-	7	0.1
CAVP	95	1.0	+	95	1.0
CAC	26	0.3	+	26	0.3
CVZ	5	0.1	+	5	0.1
CCP	1	0.0	+	1	0.0

Tabelle 2.74: Verhältnis der kontinuierlichen koordinierten Phrasen zu denjenigen mit Mutter-Diskontinuierlichkeit (MD) in NEGRA

Verteilung der Typen von Diskontinuierlichkeit in NEGRA Vom Typus MD existieren in NEGRA mit insgesamt 281 diskontinuierlichen koordinierten Phrasen recht wenige Exemplare. Sie verteilen sich wie folgt über die Phrasen:

(159) CVP (108, 38.4%), CS (75, 26.7%), CNP (51, 18.1%), CPP (29, 10.3%), CAP (11, 3.9%), CO (7, 2.5%)

Die CVP und CS, d.h. die verbalen Kategorien, sind erwartungsgemäss häufiger diskontinuierlich als die andern Kategorien. In der Tabelle 2.74 ist der Anteil der diskontinuierlichen koordinierten Phrasenkategorien in Bezug auf die kontinuierlichen Fälle aufgeschlüsselt. Sie zeigt insbesondere, dass die absolute Häufigkeit der koordinierten Phrasen nicht entscheidend ist für den Anteil an diskontinuierlichen Phrasen.

Die Anteile der insgesamt 133 Fälle von reiner Mutter-Diskontinuierlichkeit (RMD) in NEGRA sehen wie folgt aus: CS (58, 43.6%), CNP (37, 27.8%), CPP (24, 18.0%), CAP (8, 6.0%), CO (4, 3.0%), CVP (2, 1.5%). Wie erwähnt, werden bei RMD die Fälle ausgeschlossen, wo zusätzlich noch Tochter-Diskontinuierlichkeit vorhanden ist.

Die Anteile der insgesamt 37 Fälle von reiner Tochter-Diskontinuierlichkeit (RTD) in NEGRA sehen wie folgt aus: CS (32, 86.5%), CVP (5, 13.5%). Bei RTD werden die Fälle ausgeschlossen, wo zusätzlich noch Mutter-Diskontinuierlichkeit vorhanden ist. RTD ist also ein recht seltenes Phänomen, das sich auf 2 Phrasentypen beschränkt.

Phrase	Anzahl	in %	Typ	Anzahl	in %
CVP	113	35.5	MTD	106	33.3
			RTD	5	1.6
			RMD	2	0.6
CS	107	33.6	RMD	58	18.2
			RTD	32	10.1
			MTD	17	5.3
CNP	51	16.0	RMD	37	11.6
			MTD	14	4.4
CPP	29	9.1	RMD	24	7.5
			MTD	5	1.6
CAP	11	3.5	RMD	8	2.5
			MTD	3	0.9
CO	7	2.2	RMD	4	1.3
			MTD	3	0.9

Tabelle 2.75: Verteilung der Untertypen MTD, RTD und RMD bei den diskontinuierlichen koordinierten Phrasen in NEGRA.

Die beiden Verteilungen zeigen deutliche Unterschiede: In CS sind RMD und RTD stärker verbreitet, dafür verantwortlich sind insbesondere Einschubphänomene, wie sie bei direkter Rede vorkommen.

Die Anteile der insgesamt 185 Fälle von Tochter-Diskontinuität in NEGRA sehen folgendermassen aus: CVP (111, 60.0%), CS (49, 26.5%), CNP (14, 7.6%), CPP (5, 2.7%), CO (3, 1.6%), CAP (3, 1.6%).

Eine Übersicht zur Aufschlüsselung der verschiedenen Subtypen in NEGRA enthält die Tabelle 2.75. Beigestellt ist zusätzlich die Tabelle 2.76 auf der nächsten Seite, welche zeigt, dass sich die Verhältnisse in TIGER insgesamt analog verhalten.

2.7.2 Diskontinuierliche CNP

Im NEGRA-Korpus sind von den total 5169 CNP nur gerade 51 diskontinuierlich, d.h. auf der Terminalebene unterbrochen durch lexikalisches Material einer andern Konstituente. Die Tabelle 2.77 auf Seite 149 zeigt eine Übersicht aller diskontinuierlichen koordinierten CNP, welche sich im TIGER-Korpus (9857 kontinuierliche, 37 diskontinuierliche), NEGRA-Korpus sowie im CZ-Korpus (1141 kontinuierliche, 4 diskontinuierliche) befinden.

Bei manueller Durchsicht der Fälle von NEGRA wurden sechs Fehlannotationen entdeckt. Es zeigt sich weiter, dass rechtsextraponierte Relativsätze wie in (160a) knapp die Hälfte aller tatsächlichen diskontinuierlichen Fälle ausmacht. Bei den mit „?“ markierten Fällen ist unklar, ob der Relativanschluss einem Konjunkt oder der koordinierten Teilstruktur gilt. Im Beispiel (160b) ist der Relativanschluss

Phrase	Anzahl	in %	Typ	Anzahl	in %
CVP	262	40.5	MTD	242	37.4
			RTD	10	1.5
			RMD	10	1.5
CNP	140	21.6	RMD	103	15.9
			MTD	35	5.4
			RTD	2	0.3
CS	121	18.7	RMD	55	8.5
			RTD	40	6.2
			MTD	26	4.0
CPP	76	11.7	RMD	57	8.8
			MTD	17	2.6
			RTD	2	0.3
CO	22	3.4	RMD	17	2.6
			MTD	5	0.8
CAP	22	3.4	RMD	16	2.5
			MTD	4	0.6
			RTD	2	0.3
CAVP	3	0.5	MTD	2	0.3
			RMD	1	0.2
CAC	1	0.2	RMD	1	0.2

Tabelle 2.76: Verteilung der Untertypen MTD, RTD und RMD bei den diskontinuierlichen koordinierten Phrasen in TIGER.

in %	Anzahl	Bewertung	Funktion	kumulativ
43.6	24	C	RC	44
5.5	3	E	OC	49
5.5	3	C	CC	55
3.6	2	ET	—	58
3.6	2	E	MNR	62
3.6	2	C	PAR	65
3.6	2	C	OC	69
3.6	2	?	RC	73
3.6	2	?	PAR	76
1.8	1	EC	CJ	78
1.8	1	E	RC	80
1.8	1	E	PG	82
1.8	1	E	PAR	83
1.8	1	E	MO	85
1.8	1	E	GR	87
1.8	1	E	CD	89
1.8	1	E	CC	91
1.8	1	E	APP	92
1.8	1	E	AG	94
1.8	1	C	OP	96
1.8	1	?	MNR	98
1.8	1	?	APP	100

Tabelle 2.77: Die total 55 Fälle von diskontinuierlichen CNP mit Annotationsbewertung und Angabe der syntaktischen Funktion der Konstituente direkt rechts neben dem konstituentenfremden Material in NEGRA.
Legende zur Spalte „Bewertung“: C = korrekte Annotation, E = fehlerhafte Annotation, ET = Textfehler, EC = trotz fehlerhafter Annotation diskontinuierlich, ? = nicht entscheidbar.

an das zweite Konjunkt annotiert im TIGER-Korpus, die Konstruktion könnte aber auch an das erste Konjunkt oder an beide Konjunkte anschliessen, wenn „deren“ als Genitiv Plural aufgefasst wird.

(160) a. Oder es gelingt, den Kontrast zwischen $[_{CNP}$ $[_{NP}$ der Treppe als statischem Element] und $[_{NP}$ der Bewegung $\lfloor$einzufangen$\rfloor$, $[_{S-RC}$ die durch einen Menschen oder ein Tier entsteht]]] . [N_{19655}]

b. Die Entscheidung löste $[_{CNP}$ $[_{NP}$ eine Sondersitzung des Bundestages] und $[_{NP}$ eine öffentliche Kontroverse $\lfloor$aus$\rfloor$, $[_{S-RC}$ in deren Verlauf die drei Verfassungsrichter teilweise persönlich angegriffen wurden]]] . [T_{5113}]

Zwei interessante grammatische Fehler, welche in der Tabelle 2.77 mit der Funktion „—“ markiert sind, ergeben sich aus und in der Annotation von „nicht

. . . sondern“. Sie sind im Beispiel (161) gezeigt – die syntaktische Struktur ist gemäss TIGER-Annotation angedeutet. Grammatisch liegt in (161a) das Problem vor, dass das Erstglied, welches zwischen „nicht“ und „sondern“ geklammert ist, eine PP darstellt und das Zweitglied eine NP. Das Zweitglied lässt sich jedoch nur interpretieren, wenn es als „um die Bereitstellung“ reanalysiert wird. Was koordiniert wird, sind zwei PP. Meines Erachtens kann „um“ unmöglich über die Klammer von „nicht/sondern“ hinaus Skopus über „die Bereitstellung“ erhalten, was wohl auch die Annotierenden zur CNP-Struktur motiviert hat. Der Versuch der Informationkondensierung wie im Satz (161a) bringt analog auch die Grammatikalität in Beispiel (161b) zu Fall.

(161) a. Hierbei geht es wohlgemerkt [$_{PP}$ [$_{CNP}$ [$_{NP}$ nicht ⌊um ⌋ eine militärische (notfalls kämpferische) Absicherung ziviler Maßnahmen, sondern vor allem die Bereitstellung von Infrastruktur, über die Streitkräfte typischerweise verfügen (z. B. leistungsfähige Telekommunikation oder Pioniermittel)]]] . [T$_{9290}$]

b. Erst das Ende des Ost-West-Konflikts hat es möglich gemacht , daß sich mit dem Internationalen Gerichtshof in Den Haag die formal höchste Völkerrechtsinstanz mit der Legitimität von Atomwaffen befaßt; [$_{PP}$ [$_{CNP}$ [$_{NP}$ nicht nur ⌊mit⌋ der Rechtmäßigkeit ihres Einsatzes] , sondern [$_{NP}$ bereits ihrer Herstellung, ihren Tests und der Drohung mit ihnen, also der Abschreckung]]] . [T$_{32133}$]

Eine Annotation der traditionell (vgl. (Buscha 1989, 85)) als mehrteilig betrachteten Konjunktion „nicht nur. . . sondern auch“ analog zu „weder. . . noch“, wo das „weder“ immer präkoordinativ annotiert ist, würde verhindern, dass das „nicht“ zur NP gezogen werden muss und danach eine diskontinuierlich angesetzte Präposition entsteht. Eine solche Mehrwortanalyse wird in den Münsteraner Annotationskonventionen (Steiner 2003) gemacht.

2.7.3 Diskontinuierliche CS

Diskontinuierliche koordinierte Sätze erscheinen gemäss Tabelle 2.74 auf Seite 146 in 75 aller 2500 Fälle. Die Zusammensetzung der Konstituenten ist in Tabelle 2.78 auf der nächsten Seite ersichtlich. Es gibt einige wenige CS, welche andere CS rekursiv einbetten. Sinn macht dies dort wie in Beispiel (162), wo die koordinierte Tochterkonstituenten elliptisch sind.

(162) Prävention durch sportliche Betätigung, so lautet die übereinstimmende Ansicht der Experten, zahlt sich für alle aus - [$_{CS}$ [$_{S}$für die Betroffenen als verbesserte Lebensqualität] und [$_{S}$ für die Volkswirtschaft durch Verringerung der Kosten]]. [N$_{2228}$]

Dort wo Mutter-Tochter-Diskontinuierlichkeit (MTD) vorliegt, ist fast ausschliesslich das Erstglied diskontinuierlich. Das Gegenbeispiel entsteht durch unterbrochene direkte Rede, wie Beispiel (163) zeigt:

in %	Anzahl	Tochterkonjunkte	kumulativ
65.3	49	S S	65
18.7	14	S-DISCO S	84
6.7	5	S S S	91
1.3	1	VP S S	92
1.3	1	S S S S	93
1.3	1	S S-DISCO	95
1.3	1	S CS	96
1.3	1	S-DISCO CS	97
1.3	1	CS S	98
1.3	1	CS-DISCO S	100

Tabelle 2.78: Verteilung der 75 (dis-)kontinuierlichen Konjunkte von MD-diskontinuierlichen CS in NEGRA

(163) " [$_{CS}$ [$_{S}$ Die Leute wollen auch Mutter sein, Vater sein] – [$_{S}$ es gibt", ⌊ so der Personalratsvorsitzende Sittner⌋, "auch ein Leben vor dem Tod]] ." [N$_{9636}$]

Reine Tochter-Diskontinuierlichkeit entsteht bei CS bei „weder. . . noch“, wo das einleitende „weder“ im 1. Konjunkt drin steht, wie im bemerkenswerten Beispiel-Satz (164), der insgesamt 6 (!) koordinierte Satzglieder enthält. Diese Form der Koordination stellt eine Analepse dar, wobei die Position von „weder“ im 1. Konjunkt die Grenze markiert, bis zu der im 2. Konjunkt ergänzt werden muss.

(164) Das Bahnreformkonzept der Bundesregierung führt in ein ökologisches, verkehrspolitisches und gesellschaftspolitisches Desaster, [$_{CS}$ [$_{S}$ weil es ⌊[$_{KON-CD}$ weder] ⌋ in das ökologische Gesamtkonzept eingebunden ist, das den ökologischen Zielsetzungen (der Bundesregierung und der Bundestags-Enquête-Kommission zum Klimaschutz) hinsichtlich des Beitrags der Schiene zum Personennahverkehr, Personenfernverkehr sowie zum Güterverkehr entspricht] , [$_{KON-CD}$ noch] [$_{S}$ die wesentlichen Voraussetzungen schafft , um die Existenz, Minimalqualitäten und die notwendigen Verbesserungen des Nah- und Regionalverkehrs auf der Schiene zu sichern]]. [N$_{942}$]

2.7.4 Diskontinuierliche CVP

Solche Konstruktionen erscheinen gemäss Tabelle 2.74 auf Seite 146 in 108 aller 436 Fälle von CVP und stellen sowohl anteilsmässig wie absolut die häufigste, diskontinuierlich annotierte Koordinationsphrase dar. Ein typischer Fall ist in Beispiel (165) ersichtlich, bei dem ein Objekt der VP topikalisiert ist und durch das finite Hilfs- oder Modal-Verb in Verbzweitstellung (und in diesem Fall auch noch durch das Subjekt) unterbrochen wird.

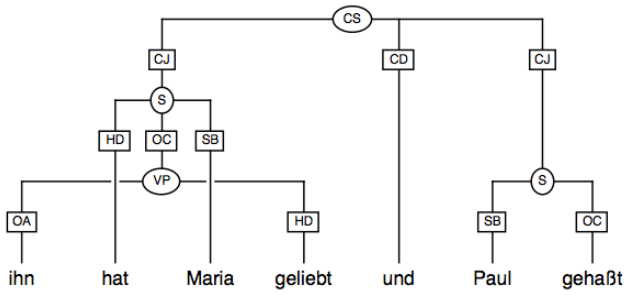
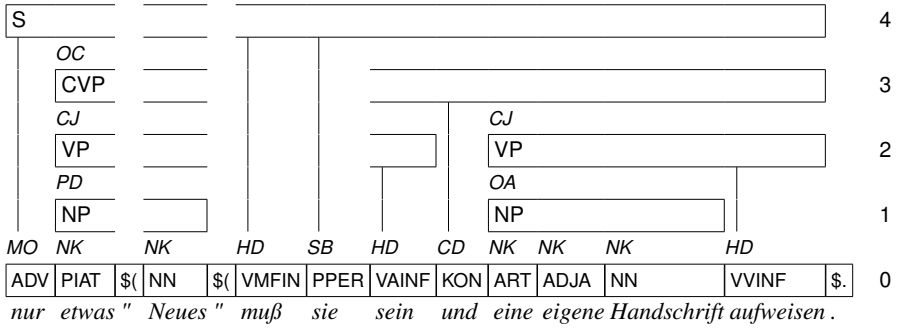


Abbildung 2.14: Annotation von koordinierten VP als CS gemäss NEGRA-Annotationshandbuch (Brants u. a. 1999, 94)

(165)



[N_4]

Wie in Abbildung 2.14 zu sehen, fordert das Annotationshandbuch (Brants u. a. 1999, 94) bei Koordinationen, wo das finite Hilfsverb und das Objekt im 2. Konjunkt fehlt, eine Annotation von CS[59].

Anders liegt der Fall gemäss Annotationshandbuch (Brants u. a. 1999, 93), wo das Subjekt und das finite Hilfsverb fehlt. Wie Abbildung 2.15 auf der nächsten Seite verdeutlicht, wird dann eine CVP verlangt. Die Annotation in Beispiel (165) entspricht somit den Richtlinien.

Neben der Verbzweitstellung eröffnet auch die Verbletztstellung die Möglichkeit von diskontinuierlichen Einschüben – wie im Beispiel (166) ersichtlich:

(166) [. . .] , [$_{S-CP}$ daß [$_{CVP}$ [$_{VP}$ sich ⌊ [$_{NN-SB}$ Kunst] ⌋ direkt oder indirekt dem gesellschaftlichen Fortschritt zu verschreiben] und [$_{VP}$ auf ein Massenpublikum Einfluß zu nehmen]] [$_{VAFIN-HD}$ habe] . [N_{1799}]

Wie die Tabelle 2.79 auf Seite 154 zeigt, erscheinen diskontinuierliche VP-Konjunkte in über 90% der Fälle nur im Erstglied. Eine Möglichkeit, wie eine diskontinuierliche VP im Letztglied entsteht, sind rechts-extraponierte Relativsätze wie in Beispiel (167):

[59] In der Penn-Treebank (Bies u. a. 1995, 126) werden solche Konstruktionen ebenfalls unter „Gapping at S-level“ behandelt.

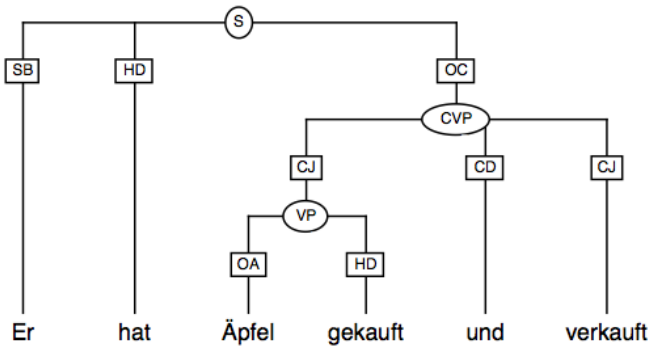
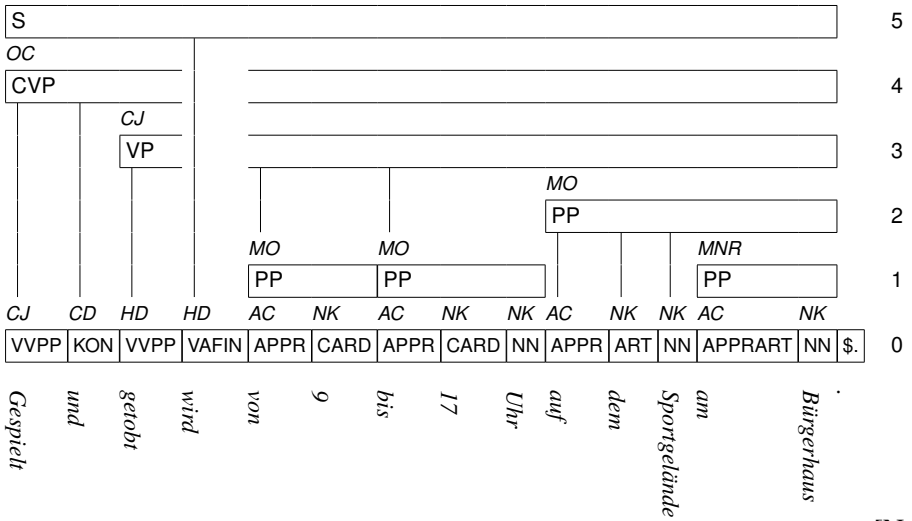


Abbildung 2.15: Annotation von CVP gemäss NEGRA-Annotationshandbuch (Brants u. a. 1999, 93)

(167) daß die für die Kinder Unterhaltspflichtigen [$_{CVP}$ [$_{VP}$ Änderungen ihrer Einkommensverhältnisse anzeigen] und [$_{VP}$ von sich aus den Unterhalt zahlen ⌊ [$_{VMFIN-HD}$ müssen ⌋ , auf den das Kind Anspruch hat]] . [N$_{3629}$]

Bei topikalisiertem koordinierten Partizip Perfekt werden die dazugehörigen Phrasen an das Letztglied gehängt wie in Beispiel (168):

(168)



[N$_{3629}$]

VP-Konjunkte mit elliptischem Verb Eine Schwierigkeit bei der Annotation von CVP stellen implizit ergänzte Konjunkte dar. Grundsätzlich können folgende Strategien verfolgt werden:

1. Nur vorhandenes Material in den Konjunkten wird annotiert, implizite Ergänzungen werden auf der primären Strukturebene nicht repräsentiert (wie

in %	Anzahl	Tochterkonjunkte	kumulativ
81.9	77	VP-DISCO VP	82
8.5	8	VP-DISCO VP VP	90
3.2	3	VP VP-DISCO	94
2.1	2	VVPP VP-DISCO	96
2.1	2	VP-DISCO VP-DISCO	98
1.1	1	VP-DISCO VVPP VP	99
1.1	1	VP-DISCO VP PP	100

Tabelle 2.79: Die Verteilung der 94 (dis-)kontinuierlichen Konjunkte in CVP aus NEGRA. Alle Vorkommen von VP-DISCO belegen diskontinuierliche VP. Ausgewertet ohne Berücksichtigung der Konjunktoren und Kommas und ohne die Fälle, wo diskontinuierliche VP nur (!) mit Einzelwörtern wie VVPP oder VVINF koordiniert sind.

bei CS).

2. Bei der Bestimmung der Phrasenkategorie wird das implizit ergänzte Material mitinterpretiert. Im „schlimmsten“ Fall fehlt dann gerade die Kopfkategorie, welche für die Phrasenbestimmung entscheidend ist.

Das Beispiel (169) illustriert eine unglückliche Vermischung von beiden Annotationsstrategien in einem Satz. Das 2. Konjunkt, dem der Verbalkopf fehlt, wird trotzdem als VP annotiert, während das 3. Konjunkt als PP angefügt wird, obwohl auch dort das „geöffnet“ implizit ergänzt werden muss für eine sinnvolle Interpretation. Der Nachteil beider Strategien wird hier sichtbar: Strategie 1 erzeugt kopflose Phrasen, Strategie 2 müsste diese Koordination als CO interpretieren.

(169) [$_{CVP}$ [$_{VP}$ geöffnet ⌊ist sie⌋ donnerstags (15.30 Uhr bis 20.30 Uhr)] und [$_{VP}$ samstags von 10 bis 14 Uhr] sowie [$_{PP}$ nach Vereinbarung]] . [N$_{10278}$]

2.7.5 Diskontinuierliche CPP

Wie in der Auflistung (159) auf Seite 146 ersichtlich, gibt es in NEGRA doch bei über 10% aller CPP Diskontinuierlichkeit. Solche Strukturen erscheinen in etwas mehr als der Hälfte dort, wo „schwere“ Konjunkte aus stilistischen Gründen ins Nachfeld genommen werden. In Beispiel (170a) mit Verbzweitstellung und Satzklammer, in Beispiel (170b) in Verbletztstellung und einem Zweitkonjunkt, das durch das Pronominaladverb „darauf“ mit dem angehängten Korrelationssatz (in NEGRA funktional annotiert als PH-RE-Kombination) schwer wird.

(170) a. Die Balletteinlage wird nicht [$_{CPP}$ [$_{PP}$ von den völlig erschöpften Tutu-Ratzen] ⌊ [$_{VVPP}$ realisiert] ⌋ , sondern [$_{PP}$ von einer rüstigen Riege Orlofskyscher Serviermädeln]] [N$_{1081}$]

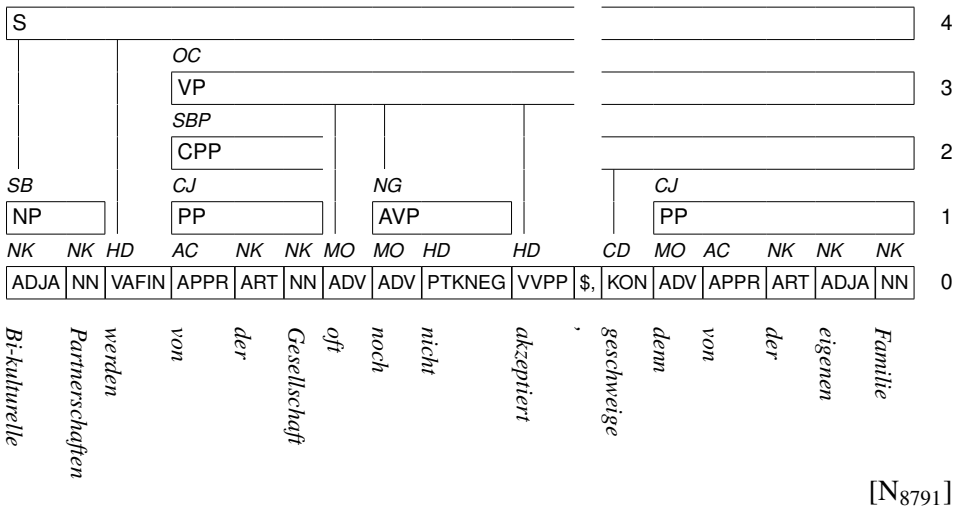
b. Es ist ein Gedicht, [$_S$ das [$_{CPP}$ [$_{PP}$ auf die nächtlich drohenden Gasraketen] ⌊ [$_{VVFIN}$ anspielt] ⌋ und [$_{PP}$ [$_{PROAV-PH}$ darauf, [$_{S-RE}$ wie sich im Blick durch die " Maske " alles unheilvoll (ja:) " verzaubert "]]] : [N$_{379}$]

Einige dieser Fälle sind aber problematisch in der Annotation: In Beispiel (171) deutet das Adverb „meist“, das nur mit einem elliptischen „ausgekocht“ zusammen interpretiert werden kann, darauf, dass hier eine elliptische CVP-Koordination vorliegt.

(171) Auch die Entscheidungsprozesse, die die streng hierarchisch strukturierte Regierungspartei fällt, werden [$_{CPP}$ [$_{PP}$ nur von einem kleinen Zirkel der " glücklichen Familie " ⌊ [$_{VVPP-HD}$ ausgekocht] ⌋ , [$_{PP}$ meist von einem einzigen Mann]] : dem Präsidenten. [N$_{651}$]

Auch in Beispiel (172) wirkt das 2. Konjunkt stark elliptisch. Wenn man sich vor Augen hält, dass bei CVP adjazent koordinierte Verbalteile als selbständige elliptische Sätze annotiert werden, bei CPP stark diskontinuierliche und elliptische Konjunkte jedoch integriert werden, zeigt sich hier eine insgesamt wenig konsistente Koordinationsannotation.

(172)



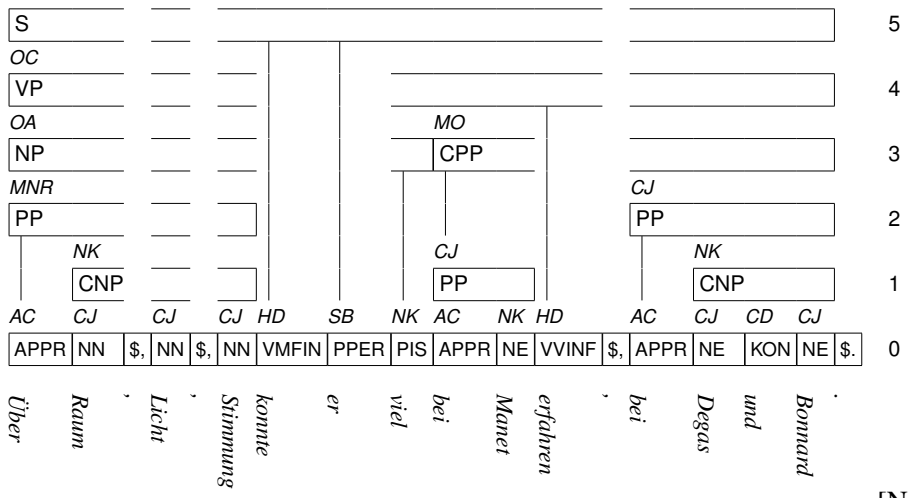
[N$_{8791}$]

Neben Annotationsproblemen ergeben sich auch grammatische Probleme, da solche Konstruktionen gerne in (allzu) komplexen Sätzen auftauchen. Die inkohärente Formulierung in Beispiel (173) sollte keine kohärente Syntaxstruktur erhalten.

(173) Warum lamentiert der Präsident des LKA Hamburg, daß "man nach über 40 Jahren Polizei im Rechtsstaat Horrorgemälde malen" würde, wenn sich gar keiner [$_{CPP}$ [$_{PP}$ gegen die bestehenden Polizeigesetze ⌊ auflehnt ⌋ , sondern [$_{PP}$ über zur Zeit noch grundgesetzwidrige, neue Möglichkeiten der Überwachung mit allen ihren Mißtrauensmöglichkeiten]] ? [N$_{4169}$]

Die Kombination von asyndetischer Koordination mit Diskontinuierlichkeit ergibt wie in Beispiel (174) mit der Rechtsherausstellung einen mündlich-fragmentarischen Stil.

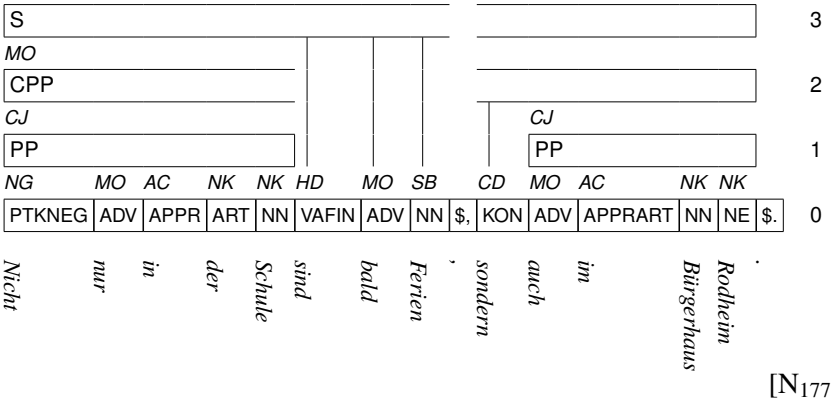
(174)



[N_{4169}]

Diskontinuierliche CPP entstehen gerne in Kombination mit dem Koordinationsmittel „nicht nur. . . sondern auch“, dessen Annotation in NEGRA in Beispiel (175) gezeigt wird. Im Gegensatz zu „entweder. . . oder“ wird kein paariger Konjunktor annotiert, sondern „nicht nur“ wird in das 1. Konjunkt integriert. Dies hat den Vorteil, dass bei einer Topikalisierung wie in (175) nur eine Konstituente im Vorfeld steht.

(175)

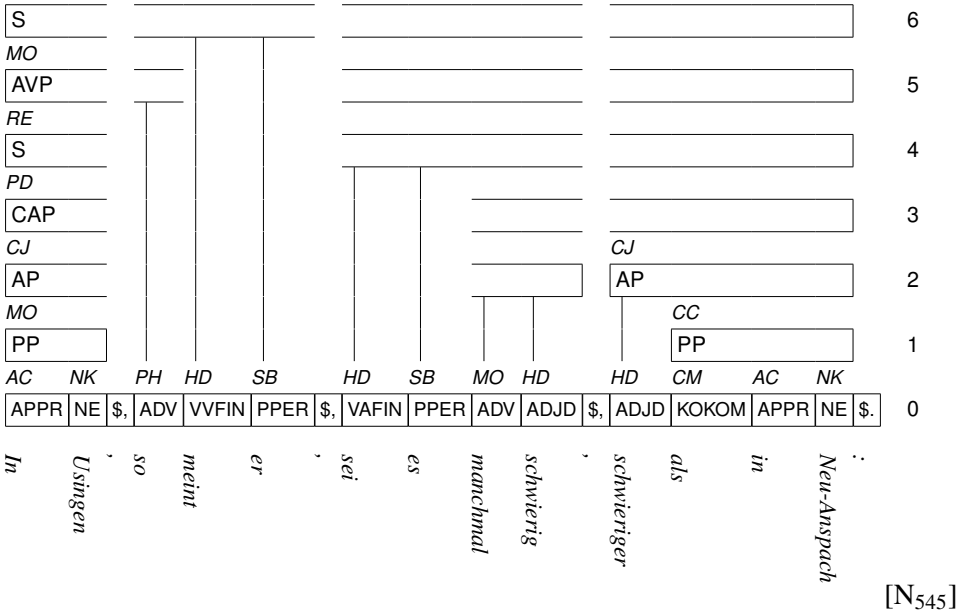


[N_{17717}]

2.7.6 Diskontinuierliche CAP

Solche Konstruktionen erscheinen gemäss Tabelle 2.74 auf Seite 146 nur in 11 Fällen. Wenn man sich diese genauer anschaut, gibt es in Beispiel (176) eine nachgestellte, satzwertige Adjektivphrase gemäss (Dudenredaktion 2005, §1327), welche rhetorisch als Anadiplose mit der Wiederaufnahme von „schwierig“ angeknüpft ist.

(176)



[N_{545}]

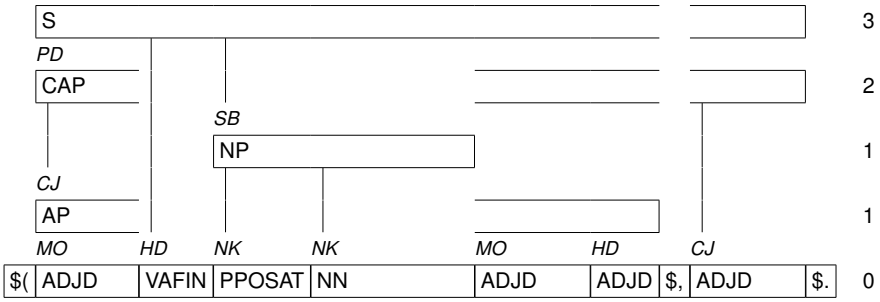
Im Fall (177) liegt eine komplexe Vergleichskonstruktion[60] vor, bei der wie häufig bei „so. . . wie“ die schwere „wie-Phrase“ nicht kontinuierlich erscheint. Denkbar ist in diesem Fall auch eine Annotation, welche von zwei Sätzen ausgeht und den 2. Satz als elliptisch betrachtet.

(177) Die Stadt steht [$_{CAP}$ [$_{AP}$ fast doppelt so hoch ⌊in der Kreide⌋ wie das doppelt so große München (2,6 Milliarden Mark)] und [$_{AP}$ dreimal so hoch wie die fast gleichgroßen Städte Dortmund (1,2 Milliarden) oder Stuttgart (1,6 Milliarden)]] . [N_{3767}]

Beim 3. Fall in (178) handelt es sich um eine ungrammatische Äusserung aus einem Interview, deren Bedeutung vermutlich mit „Meine monatliche Telefonrechnung ist wahnsinnig hoch, wahnsinnig.“ paraphrasiert werden kann. Leider gibt es in NEGRA (ebenso wenig wie in anderen Baumbanken) keine Konventionen, wie ungrammatische Sätze auszuzeichen oder zumindest auf Satzebene als ungrammatisch zu markieren sind.

(178)

[60] Eine umfassende Studie zu Vergleichskonstruktionen im Deutschen gibt Eggs (2006).



[N_{10406}]

Zusammenfassend lässt sich festhalten, dass echte diskontinuierliche CAP extrem selten vorliegen.

Kapitel 3

(Partielle) syntaktische Analyse und koordinierte Strukturen

In diesem Kapitel werden Ansätze besprochen und teilweise detailliert evaluiert, welche versuchen, für unrestringierte deutschsprachige Texte (n-)beste (partielle) syntaktische Analysen zu berechnen.

3.1 YAC-Chunking für lexikographische Akquisition und Verifikation

Das YAC-System von Kermes (2003) realisiert einen rekursionsfähigen Chunker, der insbesondere für die Unterstützung von on-line Korpus-Abfragen innerhalb des Stuttgarter Korpuserschliessungs-Systems „IMS Corpus Workbench“ (Christ und Schulze 1995) entwickelt wurde und als Regelformalismus die effiziente Abfragesprache CQP dieser Korpus-Abfrage-Umgebung verwendet.

Die verfolgten Hauptziele waren: Der Bau eines auch für lexikographische Korpusgrössen adäquat effizienten und anwendungsbereichsunabhängigen Werkzeugs, das offen für eine Textsortenspezialisierung ist. Eine robuste syntaktische Analyse, welche weder durch Mängel der grammatischen Abdeckung noch durch ungrammatisches oder fragmentarisches Textmaterial in ihrer Grundfunktion gestört wird, sowie eine klar definierte und dokumentierte Schnittstelle für die Weiterverwendung der Chunking-Resultate.

Das YAC-System stellt somit einen partiellen Parser für unbeschränkte deutsche Texte zur Verfügung, welcher tokenisierte und wortartendesambiguierte Eingaben unter Benutzung von morphologischer Information (Lezius u. a. 2000) verarbeitet.

Methode Die Konstruktion der syntaktischen Struktur und das Sammeln von phrasenspezifischen Merkmalen erfolgt in 3 Phasen:

1. Nicht-rekursive Grund-Chunks (*base-chunks*) werden mit ihrer lexikalischen und semantischen Information erzeugt.

2. Grössere und rekursive Chunks werden durch generische, handgeschriebene Regeln (CQP-Makros) erzeugt, welche nach dem Prinzip der Auswahl der längsten Konstituente arbeiten. In die Regeln integrierbar sind Listen von Wörtern oder Wortbestandteilen. Darüber hinaus erlaubt das Regelformat, die Merkmale der involvierten Elemente abzufragen und abzugleichen.

3. In der separaten, nicht-rekursiven Abschluss-Stufe werden die Strukturen und Merkmale zur endgültigen Struktur bereinigt.

Die Einbettung des YAC in die Abfragesprache eines Korpuserschliessungswerkzeug ermöglicht eine interaktive und explorative Erstellung der Chunk-Grammatik. Abdeckung und Wirkung von Regeln lässt sich als Abfrage direkt an den Fundstellen qualitativ messen. Die Einschränkung der CQP bzw. der Korpusrepräsentation, welche keine echte Repräsentation von (unbeschränkten) rekursiven Strukturen erlaubt, wird durch Hochzählen der eingebetteten Phrasen aufgehoben.

Die Integration von YAC ist allerdings nicht so stark, dass der Chunking-Prozess vollständig im Erschliessungswerkzeug möglich wäre. Die Abfragen selbst werden durch verschiedene PERL-Skripte angestossen, deren kombinierte Resultate sich letztlich in das Korpus zurückfüttern oder exportieren lassen.

Syntaktische Strukturen von YAC Die erkannten Strukturen gehen in zwei Punkten über die klassische Chunk-Definition von Abney (1996) hinaus, welche weder Rekursion einer Kategorie noch Elemente nach dem Kopf eines Chunks erlaubt. Im Beispiel (179a) von Kermes liegt eine dreifach rekursiv verschachtelte Nominalphrase vor. Das Beispiel für postnominale Erweiterungen (179b) ist dagegen etwas problematisch, weil es sich vermutlich (wenn auch ohne Kontext nicht eindeutig entscheidbar) eher um eine Adverbialphrase handelt, welche von einer NP modifiziert wird.

(179) a. [$_{NP}$ die kleinen, über [$_{NP}$ die Köpfe [$_{NP}$ der Apostel]] gesetzten Flammen]

b. [$_{NP}$ Jahre [$_{AdvP}$ später]]

Relativsätze und postnominale Präpositionalphrasen werden wegen ihrer syntaktischen Mehrdeutigkeit nicht in NP integriert. Die erkannten Strukturen umfassen Präpositionalphrasen (inklusive einfacher koordinierter NP darin), Adjektivphrasen, Adverbialphrasen, adjazente Verbalkomplexe sowie untergeordnete Teilsätze.

Auch wenn die Liste der grundsätzlich erkannten Strukturen spezifiziert ist und sich grundsätzlich am Kriterium der syntaktischen Eindeutigkeit orientiert, bleibt trotzdem für die Kodierung der Abhängigkeiten innerhalb der Strukturen Spielraum offen. So werden adjazente postnominale Genitive wie in Abbildung 3.1 immer auf der obersten Ebene angehängt – verstanden als dependenzielle Unterspe-

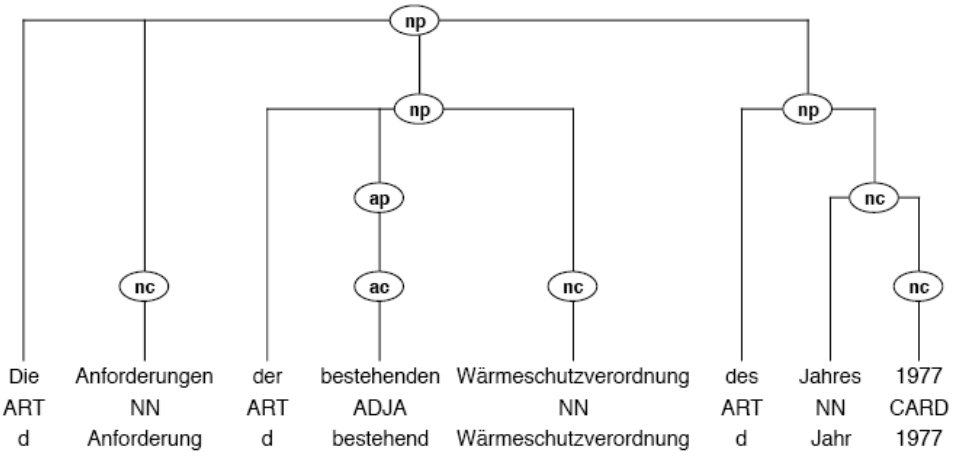


Abbildung 3.1: Annotation einer komplexen NP in YAC nach (Kermes 2003, 155)

zifikation, obwohl linguistisch gesehen das zweite postnominale Genitiv-Attribut eindeutig in das erste gehört.

Koordination ist in YAC innerhalb einer Kategorie nur auf der Ebene von maximalen Konstituenten einer syntaktischen Kategorie möglich: „YAC performs coordination only on the level of maximal constituents or implicitly if the structures are embedded in another structure“ (Kermes 2003, 152). Um Morphemkoordinationen einbettbar zu machen in NP, werden sie als NP und nicht als CNP annotiert. Problematisch wird die Einschränkung auf Koordination maximaler Konstituenten in allen Fällen, wo koordinierte Teilstrukturen als Ganzes modifiziert werden oder als Modifikatoren (z.B. Appositionen) fungieren. Eine Struktur mit koordinierten Köpfen wird deshalb nicht wie in NEGRA als (180a) repräsentiert. Wie in (180b) ersichtlich, nimmt das erste Konjunkt als „gieriges“ Konjunkt alle pränominalen Elemente auf. Analog dazu würden auch postnominale Erweiterungen nur an das letzte Konjunkt gehängt.

(180) a. [$_{NP}$ der 27-jährige [$_{CNP}$ Maler, Musiker und Komponist]]
b. [$_{CNP}$ [$_{NP}$ der 27-jährige Maler], [$_{NP}$ Musiker] und [$_{NP}$ Komponist]]

Als Fazit lässt sich festhalten, dass nicht die semantisch korrekt interpretierbare Struktur, sondern der lexikographische Nutzen im Vordergrund steht.

Evaluationen Die automatische strukturbezogene Evaluation ihres YAC-Systems gegenüber dem NEGRA-Korpus betrachtet Kermes (2003, 145) wegen der Differenzen in den beiden Sprachmodellen zurecht als wenig aussagekräftig. Die Abbildungen, welche sowohl auf den Strukturen des YAC-Outputs wie des NEGRA-Korpus notwendig sind für einen Vergleich, lassen sich auf Grund der tatsächlich repräsentierten Annotationsinformation nicht fehlerfrei durchführen. Die obersten

beiden Zeilen in Tabelle 3.1 zeigen die Resultate der Evaluation über dem ganzen NEGRA-Korpus. Wenn die Wortarten-Tags mit dem TreeTagger (Schmid 1995) berechnet werden (Fehlerrate 5,2%) und nicht als perfekte Tags dem NEGRA-Korpus entnommen sind, verschlechtern sich die Resultate des Chunkers im Schnitt um die Tagger-Fehlerquote.

Beschreibung	Precision	Recall
NP, autom. Eval., berechnete Tags	82.5	86.1
NP, autom. Eval, perfekte Tags	88.2	90.0
nur maximale NP, man. Eval., berechnete Tags	89.4	91.7
nur maximale NP, man. Eval., perfekte Tags	95.6	96.5
alle NP, man. Eval., berechnete Tags	89.9	91.7
alle NP, man. Eval., perfekte Tags	96.4	96.5

Tabelle 3.1: Resultatsübersicht von YAC in automatischer (NEGRA-Korpus) und manueller Evaluation (400 Sätze zufällig aus NEGRA) für NP

Eine manuelle Evaluation über 400 zufällig ausgewählten Sätzen aus dem NEGRA -Korpus ergibt ein vorteilhafteres Bild für das YAC-System. Wie in Tabelle 3.1 in den unteren 2 Teilbereichen ersichtlich, steigert sich insbesondere die Präzision markant. Wegen der oben erwähnten Unterspezifikation der chunk-internen Abhängigkeiten sind diese Zahlen jedoch nur sehr bedingt mit den Resultaten anderer partieller Parser und Chunker zu vergleichen.

3.2 Chunkie – partielles Parsing als Chunk-Tagging

In Skut (1999) wird ein statistisches System vorgestellt zur partiellen syntaktischen Analyse hauptsächlich von NP, PP und AP inklusive ihrer koordinierten Varianten. Im Gegensatz zur Chunking-Task, wie sie durch die CoNLL-Konferenz 2000 (Tjong Kim Sang und Buchholz 2000) definiert wurde, können Strukturen mit einer Höhe von mehr als eine Ebene entstehen. Diese Struktur ergibt sich, indem mit Hilfe des Ansatzes der HMM (*Hidden Markov Model*) die wahrscheinlichste Folge von Chunk-Tags durch den Viterbi-Algorithmus berechnet wird auf der Grundlage der Wortartentags[1]. Als Kontext wird das aktuelle Chunk-Tag und seine Wortart sowie die zwei vorangehenden Wortarten- und Chunk-Tags verwendet. Es handelt sich somit im Kern um ein Trigramm-Tagging-Verfahren, für das programmintern der Trigramm-Tagger TnT von Brants (2000b) benutzt wird. In der Abbildung 3.3 auf Seite 165 ist die Architektur des Chunkers zu sehen, der ausgehend vom Rohtext zuerst die Ebene der Wortarten-Tags berechnet und ausgehend

[1] Intern wird allerdings ein angepasstes Set von Wortarten-Tags verwendet, das insbesondere für die morphologisch markierten Wortformen von Begleitern entsprechende Kasusinformation einfügt: Alle Token, welche auf „-er“ enden und die Wortart ART tragen, bekommen das interne Kürzel ART-ngd, was die Kasus-Alternativen Nominativ, Genitiv oder Dativ kodiert. Diese „Tagfixes“, wie Skut sie nennt, werden zudem für das Vereinfachen von Tags bei Verben, aber auch für die Behandlung von Mehrwortkombinationen wie „ein wenig“ oder mehrteiligen Eigennamen benutzt.

von dieser Ebene die Chunk-Tags bestimmt. Es handelt sich hier also um einen Ansatz des „Supertaggings“ (Bangalore und Joshi 1999), d.h. das Kategorisieren von Syntaxstruktur auf den einzelnen Tokeneinheiten.

3.2.1 Chunk-Tags

Jedes Chunk-Tag besteht aus einer Phrasen-Etikette sowie einer Strukturinformation, welche das Dominanzverhältnis zwischen dem aktuellen und dem vorangehenden Token ausdrückt. Da dieses Verhältnis immer relativ zwischen zwei strukturintegrierten Nachbarknoten zueinander besteht, können im Prinzip Strukturen beliebiger Höhe aufgebaut werden. Neben den strukturintegrierten Token gibt es noch nicht-integrierte Token (z.B. Interpunktion), welche beim Verrechnen der Relationen übersprungen werden. Im Folgenden bezeichne ich mit „vorangehendem Token“ immer das am nächsten links stehende Token, das strukturintegriert ist. In der Abbildung 3.2 auf der nächsten Seite sind Beispiele für einige der nachfolgend definierten Strukturrelationen zu finden.

0 Aktuelles Token ist die Schwester vom vorangehenden Token.

1 Aktuelles Token hat keinerlei Beziehung zum vorangehenden Token.

2 Aktuelles Token ist nicht-strukturintegriert.

– Aktuelles Token ist die Nichte des vorangehenden Token.

d Aktuelles Token ist die Grossnichte des vorangehenden Token.

+ Aktuelles Token ist die Tante des vorangehenden Token.

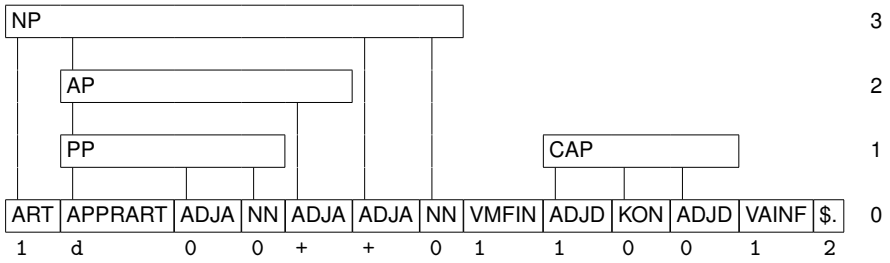
= Aktuelles Token ist die Cousine des vorangehenden Token.

Die Arbeit von Skut (1999) hat gezeigt, dass für die Trainingsmenge, welche im NEGRA-Korpus zur Verfügung steht, eine Beschränkung auf obige Strukturrelationen sinnvoll ist. Bei höherer Struktur wird die Ambiguität der Chunk-Tags pro Wortarten-Tag so gross, dass eine suboptimale Tagging-Genauigkeit für die Chunk-Erkennung entsteht.

3.2.2 Grammatik-Modell von Chunkie

In Abbildung 3.4 auf Seite 166 ist das Chunking-Resultat eines Satzes mit einer komplexen Koordination abgebildet, welcher im TIGER-Korpus die Struktur in Abbildung 3.5 auf Seite 167 erhält. Man sieht dort, dass postnominale PP nie in die NP eingebunden werden. Dies hängt selbstverständlich nur damit zusammen, wie die Bäume der Baumbank vor der Chunk-Tag-Berechnung zu Chunks geformt wurden. Rekursive Einbettung von NP ist dagegen möglich. Die pränominalen Konstruktionen erlauben im Gegensatz zu den postnominalen teilweise komplexe Konstruktionen wie in Abbildung 3.2 auf der nächsten Seite.

Wort	POS	Struktur	Etikett
Die	ART	1	NP
vom	APPRART	d	PP
statistischen	ADJA	0	PP
Chunker	NN	0	PP
vorgeschlagenen	ADJA	+	AP
syntaktischen	ADJA	+	NP
Strukturen	NN	0	NP
können	VMFIN	1	_
einfach	ADJD	1	CAP
oder	KON	0	CAP
komplex	ADJD	0	CAP
sein	VAINF	1	_
.	$.	2	_




```
(NP
  (ART  Die  )
  (AP
    (PP
      (APPRART  vom  )
      (ADJA  statistischen  )
      (NN  Chunker  ))
    (ADJA  vorgeschlagenen  ))
  (ADJA  syntaktischen  )
  (NN  Strukturen  ))
(VMFIN  können  )
(CAP
  (ADJD  einfach  )
  (KON  oder  )
  (ADJD  komplex  ))
(VAINF  sein  )
($.  .  )
```

Abbildung 3.2: Chunkie-Output: Chunk-Tags, Kastendiagramm und Klammerstruktur im Vergleich

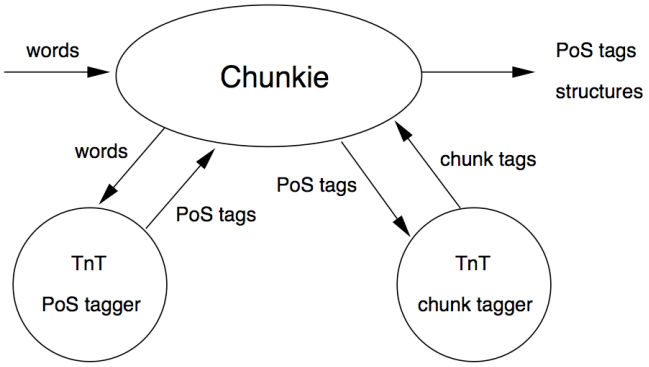


Abbildung 3.3: Architektur des Chunk-Taggers Chunkie gemäss Skut (2001, 3)

In der Auflistung (181) ist die Verteilung aller 203359 Kategorien zu sehen, welche sich bei der Evaluation des TIGER-Korpus ergeben haben. Die auf die entsprechenden Kategorien reduzierten[2] Vorkommen aus dem originalen TIGER-Korpus sind zum Vergleich in Auflistung (182) gegeben

(181) Nicht-Terminalkategorien im vom Chunkie geparsten TIGER-Korpus (Mindestvorkommen 10):
„NP“ (75809, 41.0%), „PP“ (71974, 38.9%), „AP“ (10045, 5.4%), „CNP“ (8510, 4.6%), „MPN“ (6619, 3.6%), „VZ“ (3547, 1.9%), „NM“ (3300, 1.8%), „AVP“ (2125, 1.1%), „CAP“ (1201, 0.6%), „XP“ (1154, 0.6%), „CPP“ (452, 0.2%), „CO“ (93, 0.1%), „CAVP“ (56, 0.0%), „AA“ (29, 0.0%), „MTA“ (23, 0.0%)

(182) Vergleichbare Nicht-Terminalkategorien aus dem TIGER-Korpus (Mindestvorkommen 10):
„NP“ (86335, 42.5%), „PP“ (72335, 35.6%), „AP“ (12207, 6.0%), „MPN“ (9805, 4.8%), „CNP“ (9796, 4.8%), „AVP“ (3604, 1.8%), „VZ“ (3602, 1.8%), „NM“ (2115, 1.0%), „CAP“ (1875, 0.9%), „CPP“ (1040, 0.5%), „CO“ (307, 0.2%), „CAVP“ (176, 0.1%), „AA“ (90, 0.0%), „MTA“ (30, 0.0%), „CAC“ (21, 0.0%), „CVZ“ (20, 0.0%)

Strukturprobleme Ein Problem dieses Ansatzes, der nicht auf der Verwendung von kontextfreien Grammatikregeln wie bei Brants (1999) beruht, liegt darin, dass die zugewiesenen Chunk-Tags keine konsistente Struktur ergeben können. In Skut (1999, 80) werden 3 Fälle unterschieden:

- Inkonsistente Knotenbenennung: Eine Kombination von Chunk-Tags schlägt für denselben Knoten unterschiedliche Etikettierungen vor.

[2] Es wurden (C)S,(C)VP, ISU, CH und DL ausgeblendet, sowie PN auf MPN abgebildet.

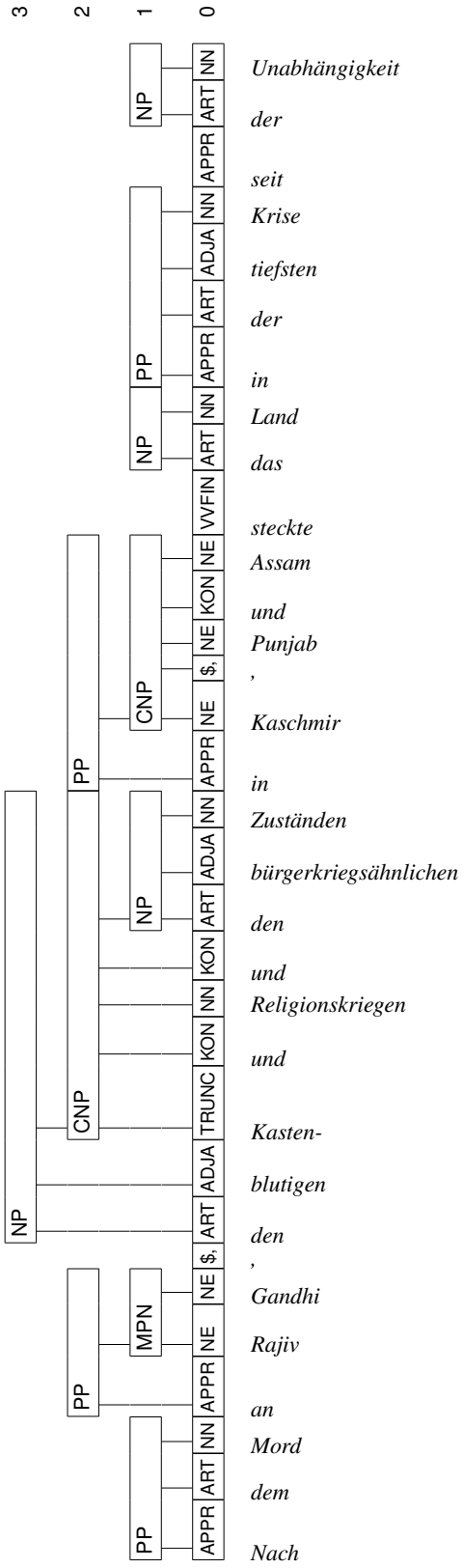


Abbildung 3.4: Chunkie-Analyse von Satz 70 des TIGER-Korpus

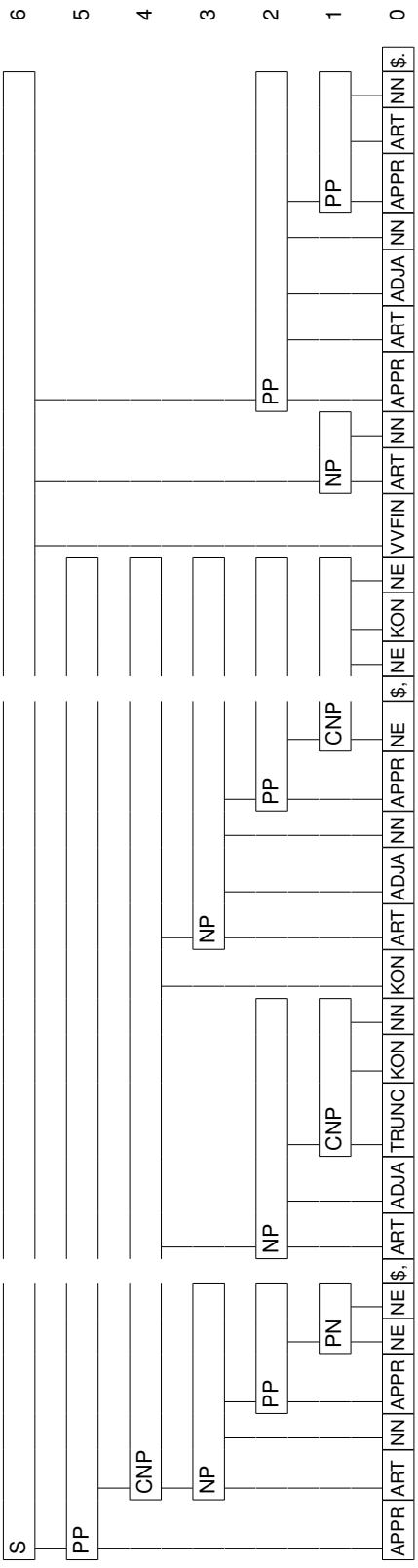


Abbildung 3.5: Original-Analyse von Satz 70 des TIGER-Korpus

- Inkonsistente Struktur: Die zugewiesenen Strukturen sind an dieser Stelle nicht möglich oder mit ihren Nachbarn nicht verträglich: z.B. weiterführende Chunk-Tags am Anfang des Satzes.
- Strukturelle Artefakte: Substrukturen der Chunks sind im Trainingskorpus nicht belegt: z.B. unäre Knoten.

Eine Nachverarbeitungsstufe versucht diese Fälle zu eliminieren. Die in Auflistung (181) sichtbaren XP hängen vermutlich damit zusammen. Strukturen, für die es in dieser Form keine „Grammatikregeln“ im Trainingskorpus gibt, ergeben sich zumindest bei den koordinierten Phrasen nicht einmal so selten. Zur Illustration ist in der Tabelle 3.2 auf der nächsten Seite die Verteilung der Tochterkonstituenten von CNP zusammengestellt. Ein relativ häufiges Phänomen stellen mit CNP annotierte Phrasen dar, welche nur aus einem Konjunkt und einem Konjunktor gebaut sind – es gibt aber auch Fälle, welche aus einem einzigen nominalen Token bestehen. Weiter kann Koordination ohne Komma oder Konjunktor sowie die Verknüpfung von unverträglichen Kategorien wie PP mit nominalen Konjunkten entstehen.

Strukturhöhen In der Auflistung (183) ist die Verteilung der Höhe derjenigen Konstituenten angegeben, welche bereits in der Auflistung (182) dargestellt wurden. Als Vergleich dient die Auflistung (184), welche die Höhen für die entsprechenden Kategorien aus dem Original-TIGER aufführt. Es zeigt sich dabei, dass insgesamt rund 90% der Strukturen im TIGER eine Höhe von 1 bis 3 aufweisen und somit durchaus im Leistungsbereich von Chunkie liegen.

(183) Verteilung der Höhe der Konstituenten im vom Chunkie geparsten TIGER-Korpus (Mittelwert:1.19 , Standardabweichung: 0.49):
„1“ (157063, 84.9%), „2“ (21540, 11.6%), „3“ (5551, 3.0%), „4“ (693, 0.4%), „5“ (94, 0.1%), „6“ (13, 0.0%), „7“ (1, 0.0%)

(184) Verteilung der Höhe der Konstituenten aus dem TIGER-Korpus (Mittelwert:1.77 , Standardabweichung: 1.25):
„1“ (120332, 59.2%), „2“ (44853, 22.1%), „3“ (19204, 9.4%), „4“ (9682, 4.8%), „5“ (4956, 2.4%), „6“ (2361, 1.2%), „7“ (1108, 0.5%), „8“ (481, 0.2%), „9“ (209, 0.1%), „10“ (95, 0.0%), „11“ (46, 0.0%), „12“ (18, 0.0%)

Ein Blick auf die Verteilung der häufigsten Chunk-Kategorie NP zeigt für Chunkie allerdings eine markante Bevorzugung der Höhe 1.

(185) Verteilung der Höhe der total 75809 NP im vom Chunkie geparsten TIGER-Korpus (Mittelwert:1.15 , Standardabweichung: 0.45):
„1“ (67373, 88.9%), „2“ (6193, 8.2%), „3“ (1929, 2.5%), „4“ (265, 0.3%), „5“ (40, 0.1%), „6“ (8, 0.0%), „7“ (1, 0.0%)

(186) Verteilung der Höhe der total 86335 NP aus dem TIGER-Korpus (Mittelwert:1.89 , Standardabweichung: 1.33):

in %	Anzahl	Tochterkonstituenten	kum.
65.9	5645	N KON N	66
7.5	640	KON N	73
5.4	459	N $, N KON N	79
1.5	126	N $, N	80
1.4	124	N KON	82
1.0	87	N N KON N	83
1.0	84	N KON N KON N	84
1.0	83	N $, N $, N KON N	85
1.0	82	PP KON N	86
0.7	61	N $, N $, N	86
0.6	54	N KON N N	87
0.6	50	KON XP	88
0.5	39	N	88
0.4	33	N N	88
0.3	25	KON N KON N	89
0.3	22	$, N KON N	89
0.3	22	N $, N $, N $, N KON N	89
0.2	21	N $, KON	90
0.2	19	N KON N $, N	90
0.2	19	PP N KON N	90
0.2	17	N KON N $, N KON N	90
0.2	17	VVPP $, N KON N	90
0.2	16	PPER KON N	91
0.2	16	XP KON N	91
0.2	15	N $, N $, N $, N	91
0.2	15	N KON XP	91

Tabelle 3.2: Verteilung der Tochterkonstituenten der CNP von Chunkie über dem TIGER-Korpus. MPN, NE, CNP und NP sind durch N repräsentiert. Alle Kommas sind als $, markiert. Gezeigt werden die Fälle mit mindestens 15 Vorkommen.

„1“ (46987, 54.4%), „2“ (20367, 23.6%), „3“ (9194, 10.6%), „4“ (4916, 5.7%), „5“ (2603, 3.0%), „6“ (1242, 1.4%), „7“ (589, 0.7%), „8“ (255, 0.3%), „9“ (95, 0.1%), „10“ (52, 0.1%), „11“ (23, 0.0%)

3.2.3 Evaluation der koordinierten Strukturen im TIGER-Korpus

In diesem Abschnitt soll die Leistung in der Erkennung von CNP, CPP und CAP gemessen werden. Die Erkennungsleistung dieses Chunking-Ansatzes wurde in (Skut 1999, 56ff.) mit unterschiedlichen Massen evaluiert: Erkennung der Chunk-Tags, Chunk-Grenzen, Chunk-Struktur. Relevant im Kontext dieser Arbeit ist die Erkennungsrate, welche erreicht wird, wenn sowohl die Wortarten- wie die Chunk-Tags automatisch berechnet werden. Skut (1999, 77) gibt für die Evaluation des NEGRA-Korpus[3], dessen Strukturen auf dieselben Chunks reduziert sind, welche für den Chunker verwendet werden, folgende Recall- und Precision-Werte für etikettierte Chunks: Reine Chunk-Tags[4] (R/P: 89%), überspannte Terminalknoten (R: 82%, P: 83%), Chunks mit innerer Struktur (R: 80%, P: 78%).

Die statistischen Trainingsdaten der im Experiment verwendeten Chunker-Version enthalten das gesamte NEGRA-Korpus. Das Grammatikmodell entspricht deshalb grundsätzlich dem NEGRA-Schema. Für eine methodisch korrekte Auswertung dürfen bei lernenden Systemen die Trainings- und Testdaten nicht überlappen. Um die Evaluationsergebnisse nicht zu verfälschen, wird im Folgendem über dem TIGER-Korpus ausgewertet.

Der Chunking-Prozess, welcher Wortarten-Tagging und Chunking, aber keine Tokenisierung umfasst, dauert für die 40020 Sätze mit 702046 Token insgesamt etwa 90 Sekunden, wenn die Standard-Beambreite von 500 genommen wird.[5] Dies entspricht einem Durchsatz von 7800 Token pro Sekunde.

Auch wenn vortokenisierter und satzsegmentierter Text als Eingabe verwendet wird, erzeugt Chunkie leider teilweise zusätzliche Satzsegmentierungen. Einige wenige lange Sätze werden zudem bruchstückhaft ausgegeben, was für die automatische Evaluation zu Synchronisationsproblemen führt. Nach dem Eliminieren der unterschiedlichen Segmente bleiben deshalb noch insgesamt 39868 Baumgraphen zurück.

Chunkie kann keine diskontinuierlichen Konstituenten erkennen. Bei der Bereitstellung des TIGER-Referenzkorpus wurden deshalb alle diskontinuierlichen Konstituenten bei ihrer 1. Unterbrechung abgeschnitten[6]. Bei der Evaluation von Systemen zur syntaktischen Analyse mit unterschiedlichen Grammatikmodellen ist die effektive Vergleichbarkeit der berechneten Resultate meist problematisch,

[3]Es handelt sich dabei um die damals zur Verfügung stehende Untermenge von 12000 Sätzen.

[4]Damit gemeint sind die letzten beiden Spalten, wie sie in der Tabellendarstellung in Abbildung 3.2 auf Seite 164 erscheinen.

[5]Gemessen auf einem 64-Bit Linux System mit Dual Core Opteron AMD mit 2600 MHz. Die Binaries des Chunkers sind 32-Bit-Applikationen.

[6]Die Tabelle 2.76 auf Seite 148 zeigt die Verteilung der vorhandenen diskontinuierlichen Koordinationskategorien.

Kategorie	vorhanden	gefunden	korrekt	P	R	F
CNP	9043	8419	4659	55.3	51.5	53.4
CAP	1853	1196	781	65.3	42.2	51.2
CPP	1020	451	94	20.9	9.2	12.8
CAVP	174	55	18	32.7	10.3	15.7

Tabelle 3.3: Evaluation der erkannten Koordinationskategorien von Chunkie über TIGER. Gezählt wurden alle Chunks, welche nicht rekursiv Chunks derselben Kategorie einbetten. Legende: P: Precision, R: Recall, F: F-Mass

was bei Kübler (2005) für Deutsch oder bei Lin (1995) bezüglich PARSEVAL (Gaizauskas u. a. 1998) diskutiert wird. Eine dort vorgeschlagene Möglichkeit zu aussagekräftigeren Evaluationsmethoden besteht darin, die Kopfrelationen auszuwerten anstelle der syntaktischen Phrasenstruktur. Ein in Clematide (2002) präsentierter alternativer Ansatz besteht darin, selektiv bestimmte syntaktische Konstruktionen als Klassifikation auf die Terminalebene zu projizieren und diese dort mit der Evaluationstechnik für Klassifikationen zu erheben, wie sie prototypisch für Wortarten-Tagging sind. Dies wird in den folgenden Abschnitten gemacht, da Chunkie keine Zuweisung der Kopffunktion in den Chunks macht.

3.2.3.1 Evaluation der Koordinationskategorien: Grenzen und Typen der Chunks

Die Resultate der Evaluation, welche als Evaluationskriterium die etikettierten überspannten Terminalknoten im IOB-Chunk-Format (Tjong Kim Sang und Buchholz 2000) verwendet, sind in der Tabelle 3.3 zusammengestellt. Dabei wurden für die häufigsten Koordinationskategorien, welche Chunkie erkennt, selektiv und jeweils separat der grösste Chunk, welcher keinen Chunk derselben Kategorie dominiert, auf die Terminalebene projiziert. Damit lassen sich Beginn, Ende und Typ des Chunks auswerten.

Die Erkennungsrate aller koordinierten Kategorien liegt erwartungsgemäss tiefer als die von Skut berechnete Gesamtrate. Während CNP und CAP mit einem F-Mass von gut 50% recht mässige Werte erhalten, ist die Erkennungsrate bei CPP extrem schlecht. Ein Grund lässt sich aus der Tabelle 3.4 auf Seite 173 ablesen, welche die Höhe der verschiedenen Koordinationskategorien zwischen dem TIGER-Vergleichskorpus und den Resultaten von Chunkie vergleicht. Da die PP selbst fast ausschliesslich phrasal sind, ist die Höhe dieser Konstituenten im Minimum 2. Wie Skut (1999, 74) gezeigt hat, sinkt die Erkennungsrate des Chunkers mit zunehmender Höhe der Strukturen. Da aber trotzdem in TIGER 34% aller CPP die Höhe 2 haben, erscheinen solch schlechte Ergebnisse nicht zwingend. Ein möglicher Faktor könnte darin bestehen, dass im Grammatikmodell von Chunkie alle postnominalen PP-MNR aus den Chunks ausgelöst werden, was gemäss Tabelle 2.61 auf Seite 123 rund 1/4 der PP betrifft. Der Einbau von PP in grössere

Strukturen kommt somit global kaum vor, was für ein lernendes, probabilistisches Verfahren entsprechende Konsequenzen hat.

Weiter lässt sich aus der Tabelle ablesen, dass sowohl bei CNP mit 72% und CAP mit 84% ein Grossteil der Fälle mit einer Strukturtiefe vorhanden ist, die für Chunkie ideal wäre. Bei den CAP mögen die in Tabelle 2.8 auf Seite 23 zusammengestellten Mängel von NEGRA eine Rolle spielen, da koordinierte Adjektive nicht immer konsequent als CAP ausgezeichnet worden sind.

3.2.3.2 Anzahl überspannter Terminale

Eine andere Sicht auf die Komplexität der Chunks ergibt sich, wenn man die Anzahl lexikalischer Terminalknoten untersucht, welche jeweils von einer koordinierten Phrase dominiert werden. Die Tabelle 3.5 auf Seite 174 zeigt, dass die CPP in TIGER eine mittlere Länge von knapp 12 Token aufweisen. Auch wenn man noch die normalerweise darin enthaltenen 2 Präpositionen abzieht, ergibt sich mit 10 Token im Vergleich zu den CNP mit einer mittleren Länge von 6 Token eine deutlich höhere Komplexität.

In der Tabelle 3.6 auf Seite 175 sieht man die Resultate der Evaluation der von Chunkie erkannten Koordinationskategorien, wenn man nach der Token-Anzahl der Chunks aufschlüsselt. Die Spalte Präzision zeigt somit, wie stark man den Resultaten von Chunkie vertrauen kann in Abhängigkeit von der Token-Anzahl der gefundenen Chunks. So sind die Unterschiede bei der Erkennung von CNP stark abfallend: Chunks der Länge 3 haben noch 72% Präzision, aber schon Chunks der Länge 4 haben nur noch 46%. Ein ganz ähnliches Bild zeigt sich bei den CAP, wo Chunks der Länge 4 nicht einmal zur Hälfte so korrekt sind wie die Chunks der Länge 3.

Die Resultate zeigen, dass grundsätzlich recht lange Chunks gebildet werden können mit diesem Verfahren, dass aber die Qualität der längeren Analysen wenig zuverlässig ist, was für die Erkennung von Koordinationskategorien besonders negativ ins Gewicht fällt. Es bleibt die Frage offen, ob mit deutlich mehr Trainingsmaterial bessere Resultate zu erwarten sind. Die Lernkurve der Gesamtkorrektheit in Skut (1999, 74) flacht allerdings schon bei 12000 Trainingssätzen deutlich ab. Ein anderer Punkt ist die Behandlung der Interpunktionen, welche oberflächlich betrachtet als strukturell integriert erscheinen, beim Berechnen der Chunk-Tags in Tat und Wahrheit jedoch ignoriert werden. Für die Erkennung von asyndetischer Koordination wäre die Berücksichtigung der Funktion des Kommas jedoch ein wichtiger Faktor.

TIGER

CNP (total 9796)
(Mittelwert: 2.01 ±1.34)

in %	Anzahl	Höhe	kum.
49.0	4798	1	49
23.5	2299	2	72
14.8	1445	3	87
7.0	687	4	94
3.6	349	5	98
1.2	113	6	99
0.6	56	7	100
0.3	28	8	100
0.1	13	9	100

CPP (total 1040)
(Mittelwert: 3.29 ±1.4)

in %	Anzahl	Höhe	kum.
34.0	354	2	34
33.4	347	3	67
15.1	157	4	82
10.2	106	5	93
3.8	40	6	96
2.0	21	7	98

CAP (total 1875)
(Mittelwert: 1.54 ±1.02)

in %	Anzahl	Höhe	kum.
69.6	1305	1	70
14.4	270	2	84
11.6	218	3	96
2.8	52	4	98
0.6	12	5	99

CAVP (total 176)
(Mittelwert: 1.17 ±0.54)

in %	Anzahl	Höhe	kum.
89.2	157	1	89

Chunkie

CNP (total 8510)
(Mittelwert: 1.5 ±0.66)

in %	Anzahl	Höhe	kum.
58.4	4971	1	58
34.6	2943	2	93
5.9	499	3	99
1.1	95	4	100

CPP (total 452)
(Mittelwert: 2.19 ±0.41)

in %	Anzahl	Höhe	kum.
82.1	371	2	82
17.3	78	3	99

CAP (total 1201)
(Mittelwert: 1.23 ±0.55)

in %	Anzahl	Höhe	kum.
83.1	998	1	83
11.6	139	2	95
4.8	58	3	100

CAVP (total 56)
(Mittelwert: 1.25 ±0.72)

in %	Anzahl	Höhe	kum.
87.5	49	1	88

Tabelle 3.4: Verteilung der Höhe der Koordinationsstrukturen in TIGER (linke Spalte) und im vom Chunkie analysierten TIGER-Korpus (rechte Spalte). Gezeigt werden Fälle mit Mindestvorkommen von 10, die Prozentangaben beziehen sich auf alle Fälle.

TIGER

CNP (total 9796)
(Mittelwert: 6.21 ±5.48)

in %	Anzahl	Länge	kum.
41.3	4044	3	41
11.3	1105	4	53
10.8	1062	5	63
9.7	951	10-14	73
6.3	619	6	79
5.5	538	7	85
3.9	386	15-19	89
3.5	346	9	92
3.4	336	8	96
2.8	272	20+	98
1.4	137	2	100

CPP (total 1040)
(Mittelwert: 11.72 ±7.01)

in %	Anzahl	Länge	kum.
27.7	288	10-14	28
14.5	151	15-19	42
11.5	120	8	54
10.7	111	6	64
10.0	104	7	74
9.1	95	9	84
9.0	94	20+	92
6.5	68	5	99

CAP (total 1875)
(Mittelwert: 3.95 ±2.83)

in %	Anzahl	Länge	kum.
58.0	1087	3	58
11.1	208	4	69
11.0	206	2	80
6.8	127	5	87
3.7	69	6	91
2.9	55	7	94
2.1	39	10-14	96
1.6	30	9	97
1.5	29	8	99
0.9	17	15-19	100

Chunkie

CNP (total 8510)
(Mittelwert: 4.31 ±2.06)

in %	Anzahl	Länge	kum.
51.9	4415	3	52
16.3	1391	5	68
9.5	805	4	78
6.9	585	6	85
5.2	440	7	90
2.8	241	8	93
2.5	216	10-14	95
2.3	192	2	97
1.8	151	9	99
0.5	40	1	100
0.4	33	15-19	100

CPP (total 452)
(Mittelwert: 5.93 ±2.24)

in %	Anzahl	Länge	kum.
19.2	87	5	19
17.3	78	6	36
14.2	64	4	51
13.5	61	7	64
11.7	53	3	76
10.4	47	8	86
7.1	32	10-14	93
4.4	20	9	98

CAP (total 1201)
(Mittelwert: 3.62 ±1.34)

in %	Anzahl	Länge	kum.
71.5	859	3	72
9.5	114	5	81
7.4	89	4	88
4.1	49	6	92
3.3	40	7	96
1.7	21	2	98
1.0	12	8	98

Tabelle 3.5: Verteilung der Anzahl Terminale der Koordinationsstrukturen in TIGER (linke Spalte) und im vom Chunkie analysierten TIGER-Korpus (rechte Spalte). Gezeigt werden Fälle mit Mindestvorkommen von 10, die Prozentangaben beziehen sich auf alle Fälle.

CNP

L	P	R	F	gef.
3	72.4	85.4	78.4	4296
4	46.3	55.4	50.4	1216
5	47.1	50.1	48.6	1046
6	35.8	38.3	37.0	634
7	33.1	23.7	27.6	378
8	21.3	12.0	15.3	183
10	31.2	7.5	12.1	64
9	27.6	8.2	12.6	98
11	23.9	5.3	8.6	46
2	5.5	13.9	7.9	380
15	60.0	3.3	6.3	5
12	18.2	2.3	4.1	22
13	33.3	2.0	3.9	9
16	50.0	1.5	3.0	2
17	100.0	1.6	3.1	1
1	0.0	0.0	0.0	36
14	0.0	0.0	0.0	3

CPP

L.	P	R	F	gef.
6	37.8	24.8	29.9	74
5	23.3	28.2	25.5	86
8	42.2	16.5	23.8	45
7	21.9	13.1	16.4	64
9	31.8	8.0	12.7	22
10	33.3	6.8	11.3	12
11	50.0	3.2	6.0	4
12	0.0	0.0	0.0	2
13	0.0	0.0	0.0	1
2	0.0	0.0	0.0	10
3	0.0	0.0	0.0	65
4	0.0	0.0	0.0	66

CAP

L.	P	R	F	gef.
3	81.4	63.8	71.5	840
4	38.1	24.9	30.1	134
5	23.8	15.8	19.0	84
6	7.1	5.8	6.4	56
9	20.0	3.6	6.1	5
8	10.0	3.6	5.3	10
2	8.6	1.4	2.5	35
1	0.0	0.0	0.0	4
10	0.0	0.0	0.0	1
12	0.0	0.0	0.0	1
14	0.0	0.0	0.0	1
7	0.0	0.0	0.0	25

CAVP

L.	P	R	F	gef.
3	50.0	11.8	19.1	36
10	0.0	0.0	0.0	2
2	0.0	0.0	0.0	7
4	0.0	0.0	0.0	5
5	0.0	0.0	0.0	1
7	0.0	0.0	0.0	2
8	0.0	0.0	0.0	1
9	0.0	0.0	0.0	1

Tabelle 3.6: Evaluation der erkannten Koordinationskategorien von Chunkie über TIGER aufgeschlüsselt nach Länge. Legende: L.: Anzahl lexikalischer Token (keine Interpunktion), P: Precision, R: Recall, F: F-Mass, gef.: Anzahl der von Chunkie gefundenen Chunks

3.3 Der Gojol-Parser – ein robuster Parser für Deutsch

Beim Parsing-System von V. Gojol handelt es sich um eine Kombination von Tagger und Parser[7], welcher einerseits eine dependenzorientierte Analyse und andererseits eine daraus abgeleitete kontextfreie Konstituentenstruktur erzeugt. In der Abbildung 3.6 auf der nächsten Seite ist die kontextfreie Konstituentenstruktur im Penn-Treebank-Format gezeigt und darunter die zugrunde liegende Dependenzinformation in einem tabellarischen Textformat: Spalte 1 identifiziert dabei die Wortposition, Spalte 3 die Position des Kopfes, der dieses Wort regiert (bzw. `root` für unregierte Köpfe), Spalte 4 enthält die Wortart, Spalte 5 etikettiert die Abhängigkeit.

Das Parsing-System ist teilweise statistisch und beinhaltet ein korpusbasiertes Trainingsmodul. Die getestete Version hat ein syntaktisches Sprachmodell, das aus den ersten 2000 Sätzen des NEGRA-Korpus destilliert wurde. Die Analysen des Parsers wollen aber in keiner Weise die NEGRA-Annotation reduplizieren, sondern stehen stark in der Tradition der Dependenzgrammatik: Satzstruktur, phrasale Kategorien und funktionale Abhängigkeiten unterscheiden sich markant vom Grammatikmodell von NEGRA.

Die Ebene der Präterminale weist eine nahe Verwandtschaft zum STTS-Tagset auf. Von statistischen Taggern für Deutsch schwierig zu unterscheidende Kategorien wie NN und NE werden vereinigt, da sie meist die Hauptfehlerquelle bilden.[8] Weiter werden sowohl Adverbien als auch prädikativ und adverbial verwendete Adjektive grundsätzlich mit ADJD getaggt. Eine vergleichende Evaluation bezüglich ausgewählter nicht-koordinativer Konstruktionen findet sich in Clematide (2002).

Die hier verwendete und evaluierte Version besitzt keine echte morphologische Analyse, ein heuristisches Ersatzmodul liefert mehr oder weniger korrekte morphosyntaktische Merkmale. Die initiale Version des Parsers umfasst ein Lexikon, das aus den ersten 10'000 Sätzen des NEGRA-Korpus stammt. Während des Taggens werden verarbeitete unbekannte Wörter laufend hinzugefügt. Die Robustheit des Parsers sowie die Anzahl der zurückgelieferten Analysen ist beim Verarbeiten konfigurierbar. Der Speicherbedarf beträgt nur wenige Megabyte; um 100 Sätze (16.5 Wörter pro Satz) am Stück zu parsen, braucht es 40 Sekunden auf einem SUN-Rechner (Sparc V9 Prozessor 750 MHz).

Im Prinzip unterstützt der Parser auch die Ausgabe von mehreren Analysen pro Satz, sogenanntes *n-best-Parsing*. Bei näherer Betrachtung hat sich allerdings gezeigt, dass die Varianten nicht dem entsprechen, was man sich linguistisch un-

[7] Der Parser kann unter `http://www.cl.uzh.ch/siclemat/sprachanalyse/gojol/` ausprobiert und dessen Ausgabe im leicht lesbaren Kastendiagramm-Format betrachtet werden. Der Entwickler des Systems, Dr. ing. V. Gojol aus Bukarest kann unter `gojol@rnc.ro` kontaktiert werden. Der Parser ist allerdings nicht frei verfügbar.

[8] Für Chunkie allerdings sind NE gerade wichtige Kategorien, da adjazente Folgen davon grundsätzlich als mehrteilige Eigennamen (MPN) zusammengezogen werden. Dem Gojol-Parser entfällt durch diese Optimierung des Taggers eine wichtige Entscheidungsgrundlage.

```
(S
  (.VBPRD
    (.SUBJ
      (NP
        (.DET
          (PDAT _Dieses))
        (NN Stadium)))
    (VVFIN dauert)
    (AJP
      (.DET
        (ART eine))
      (ADJA halbe))
    (PP
      (APPR bis)
      (.PREP
        (NP
          (.DET
            (CARD zwei))
          (NN Stunden)))))
  ($. .))
```

```
1  _Dieses     ->2   PDAT    .DET->NN
2  Stadium     ->3   NN      .SUBJ->VVFIN
3  dauert      root  VVFIN
4  eine        ->5   ART     .DET->ADJA
5  halbe       ->3   ADJA    AJP->VVFIN
6  bis         ->3   APPR    PP->VVFIN
7  zwei        ->8   CARD    .DET->NN
8  Stunden     ->6   NN      .PREP->APPR
9  .                 $.
```

Abbildung 3.6: Phrasenstrukturausgabe und Dependenzausgabe des Gojol-Parsers

ter syntaktisch möglichen, aber schwierig zu disambiguierenden Alternativen vorstellt. Deshalb wurde für nachfolgenden Evaluation nur das beste Parse-Resultat verwendet. Für die möglichen Syntaxstrukturen existiert keine explizite Grammatik – explizit lassen sich nur unerwünschte Grammatikregeln verbieten.

Anzahl	Kürzel	Kurzbeschreibung
178282	NP	Nominalphrase oder Bestandteil davon
136810	.DET	Determinator, Quantor oder attributive Adjektive
68934	PP	Präpositionalphrase
64984	.PREP	In PP von Präposition abhängige Phrase
60220	.VBPRD	Verbalhaltige Konstituente
48712	.SUBJ	Subjekt
35230	S	Satz
17598	.AUX	Von Hilfs- oder Modalverb abhängige Verbalphrase
17344	AJP	Komplex aus Determinatoren und Adjektiven
15058	.PLUSD	Erstglied von koordinierten .DET
13960	AVP	Adverbialphrase
13900	.GNTV	Post- oder pränominales Genitivattribut
12218	.PLUSN	Erstglied von koordinierten Nominalphrase
6200	.ADV	Modifizierende Adverbialphrase oder prädikatives Adjektiv
5908	VPP	Verbalphrase mit Partizip Perfekt
5794	S-	Defizienter Satz (meist ohne finites Verb)
4550	VIP	Verbalphrase mit infinitem Verb
4532	.SBC	Finiter Nebensatz mit Subjekt
4216	.RFLX	Reflexivpronomen
3884	.VBSUP	Negationspartikel (wie NG in NEGRA)
3614	VZP	Verbalphrase mit „zu“-Infinitiv
2824	VZ	„zu“ mit Infinitiv (wie NEGRA)
2730	.PLUSV	Erstglied einer koordinierten verbalen Phrase
2460	.SBA	Relativsatz mit Relativpronomen als Subjekt
546	.MISC	Unspezifische Sammelkategorie
502	.PLUSA	Erstglied von koordinierten .AUX
330	.GNTVF	Pränominales Genitivattribut mit „-s“
138	.POST	In PP von Postpositionen abhängige Phrase

Tabelle 3.7: Übersicht über die syntaktischen Kürzel des Gojol-Parsers mit Häufigkeitsangaben über dem geparsten NEGRA-Korpus

3.3.1 Das Grammatik-Modell des Gojol-Parsers

Die grosse Robustheit des Parsers hängt stark mit seinem unterspezifizierten Grammatik-Modell zusammen. Die Tabelle 3.8 auf Seite 180 enthält eine Übersicht über die häufigsten Phrasenstrukturen, welche sich in der Ausgabe des Parsers befinden. Insgesamt verwendet der Gojol-Parser 18404 verschiedene Regeln, wenn man die

Kommata und andere satzinterne Interpunktion nicht berücksichtigt. Die Tabelle 3.7 auf der vorherigen Seite gibt eine Übersicht zu den im Parser verwendeten Phrasenkategorien. Die Kategorien mit einem Punkt am Anfang haben tendenziell den Charakter von syntaktischen Funktionen und können – sofern sie unäre Produktionen darstellen – als solche in die Baumstruktur integriert werden (vgl. Abbildung 3.8 auf Seite 182). Leider ist dies aber vom Parser nicht durchgängig so gemacht und vom Entwickler auch nicht so intendiert. Insbesondere die recht häufige Doppeldominanz von .VBPRD ist redundant bzw. idiosynkratisch.

3.3.1.1 Nominalphrasen

Wie man der Tabelle 3.8 auf der nächsten Seite entnehmen kann, bestehen Nominalphrasen zur Mehrheit aus dem Modifikator .DET und einem Kopf. Der Modifikator .DET muss als funktionale Sammel-Kategorie aufgefasst werden, da er syntaktisch gesehen eine heterogene Konstituenz aufweist. Die Auflistung (187) zeigt, dass er zwar hauptsächlich Artikel und attributive Adjektive umfasst, daneben aber auch knapp 8% NN.

(187) Expansionen der total 136810 .DET-Kategorie des Gojol-Parsers über NEGRA mit mehr als 1% Anteil:
„ART“ (70288, 51.4%), „ADJA“ (27610, 20.2%), „NN“ (10668, 7.8%), „AJP“ (9402, 6.9%), „CARD“ (7470, 5.5%), „PPOSAT“ (3500, 2.6%)

Dies hat damit zu tun, wie bei der Grammatik-Transformation mit der NEGRA-Kategorie MPN für mehrteilige Eigennamen und Appositionen umgegangen wird. Aus den Resultaten des Parsers kann man schliessen, dass alle NP und MPN als Struktur mit rechtsperipherem Kopf betrachtet werden. So wird für die beiden NP mit mehrteiligen Eigennamen aus dem Fragment, das im NEGRA-Korpus wie in (188a) annotiert ist, die Struktur in (188b) berechnet. Die Erkennung von engen Appositionen und mehrteiligen Namen ist wegen der Nicht-Unterscheidung von NN und NE fast unmöglich und führt zu einer schlechten Kopfbestimmung in komplexen NP.

(188) a. [...] die in Zusammenarbeit mit [$_{MPN}$ [$_{NE}$ Philips] [$_{NE}$ Classics]] [$_{NP}$ das [$_{NN}$ Label] [$_{MPN}$ [$_{NE}$ Point] [$_{NE}$ Music]]] ins Leben gerufen haben. [N_{12}]

b. [...] die in Zusammenarbeit mit [$_{.DET}$ [$_{NN}$ Philips] [$_{NN}$ Classics]] [$_{NP}$ [$_{.DET}$ das] [$_{.DET}$[$_{NN}$ Label] [$_{NN}$ Point]] [$_{NN}$ Music]]] ins Leben gerufen haben. [N_{12}]

3.3.1.2 Koordinierte Strukturen

Im konkreten Sprachmodell des Gojol-Parsers sind koordinierte Strukturen nicht flach und symmetrisch nebeneinander geordnet wie im NEGRA-Annotationsmodell, sondern sie werden rechtsköpfig mit rekursiver Expansion des Erstglieds ana-

in %	Anzahl	Phrase	Tochterkonstituenten	kumulativ
13.4	97964	NP	.DET NN	13.4
9.6	70288	.DET	ART	23.0
7.1	52012	PP	APPR .PREP	30.1
5.8	42276	.PREP	NP	35.9
4.2	30818	S	.VBPRD	40.1
3.8	27610	.DET	ADJA	43.9
3.6	26372	.SUBJ	NP	47.5
2.4	17740	NP	.DET NP	49.9
2.4	17534	.PREP	NN	52.3
2.1	15332	.VBPRD	.VBPRD	54.4
1.8	13174	NP	NP PP	56.2
1.8	13172	.GNTV	NP	58.0
1.7	12740	PP	APPRART .PREP	59.7
1.5	10668	.DET	NN	61.2
1.3	9592	NP	NP .GNTV	62.5
1.3	9402	.DET	AJP	63.8
1.1	8154	.SUBJ	PPER	64.9
1.0	7470	.DET	CARD	65.9
1.0	7414	AVP	.PLUSD ADJD	66.9
0.9	6440	.PLUSN	NP	67.8
0.9	6304	.PLUSD	ADJD	68.7
0.9	6266	.AUX	VVPP	69.6
0.8	5892	.ADV	ADJD	70.4
0.8	5714	.SUBJ	NN	71.2
0.7	5288	NP	NN PP	71.9
0.7	4760	NP	.DET .DET NN	72.6
0.6	4552	.AUX	VPP	73.2
0.6	4390	.PLUSN	NN	73.8
0.6	4216	.RFLX	PRF	74.4
0.6	4166	AJP	.PLUSD ADJA	75.0
0.5	3948	NP	.PLUSN KON NN	75.5
0.5	3884	.VBSUP	PTKNEG	76.0
0.5	3714	NP	NN .GNTV	76.5
0.5	3714	.SUBJ	PRELS	77.0
0.5	3500	.DET	PPOSAT	77.5
0.5	3484	AVP	ADJD PP	78.0
0.4	3058	.AUX	VIP	78.4
0.4	2760	AJP	.ADV ADJA	78.8
0.4	2668	.AUX	VVINF	79.2
0.3	2500	VZ	PTKZU VVINF	79.5

Tabelle 3.8: Verteilung der 40 häufigsten Grammatikregeln des Gojol-Parsers über dem NEGRA-Korpus. Insgesamt ergibt sich über dem NEGRA-Korpus bei total 731480 Phrasen mit 18404 Types ein Type-Token-Verhältnis von 1:39.7 .

```
1   _Folklore          ->3   NN      .PLUSN->NN
2   ,                        $,
3   Rock               ->5   NN      .PLUSN->NN
4   ,                        $,
5   Klassik            ->7   NN      .PLUSN->NN
6   und                      KON
7   Jazz               ->9   NN      NP->VZ
8   zu                 ->9   PTKZU   PTKZU->VVINF
9   vermischen         ->10  VVINF   VZP->VVFIN
10  reicht             ->19  VVFIN   .PLUSV->VAFIN
11  ihnen              ->10  PPER    PPER->VVFIN
12  nicht              ->10  PTKNEG  .VBSUP->VVFIN
13  ,                        $,
14  sie                ->15  PPER    .SUBJ->VVFIN
15  nutzen             ->19  VVFIN   .PLUSV->VAFIN
16  die                ->17  ART     .DET->NN
17  Elektronik         ->10  NN      NP->.VBPRD
18  und                      KON
19  sind               root  VAFIN
20  sogar              ->21  ADJD    .PLUSD->PROAV
21  dazu               ->22  PROAV   AVP->VVPP
22  übergegangen       ->19  VVPP    .AUX->VAFIN
23  ,                        $,
24  Instrumente        ->27  NN      NN->VZ
25  selbst             ->27  ADJD    ADJD->VZ
26  zu                 ->27  PTKZU   PTKZU->VVINF
27  bauen              ->19  VVINF   VZP->VAFIN
28  .                        $.
```

Abbildung 3.7: Die Dependenzstruktur des Gojol-Parsers für Satz 3 aus NEGRA

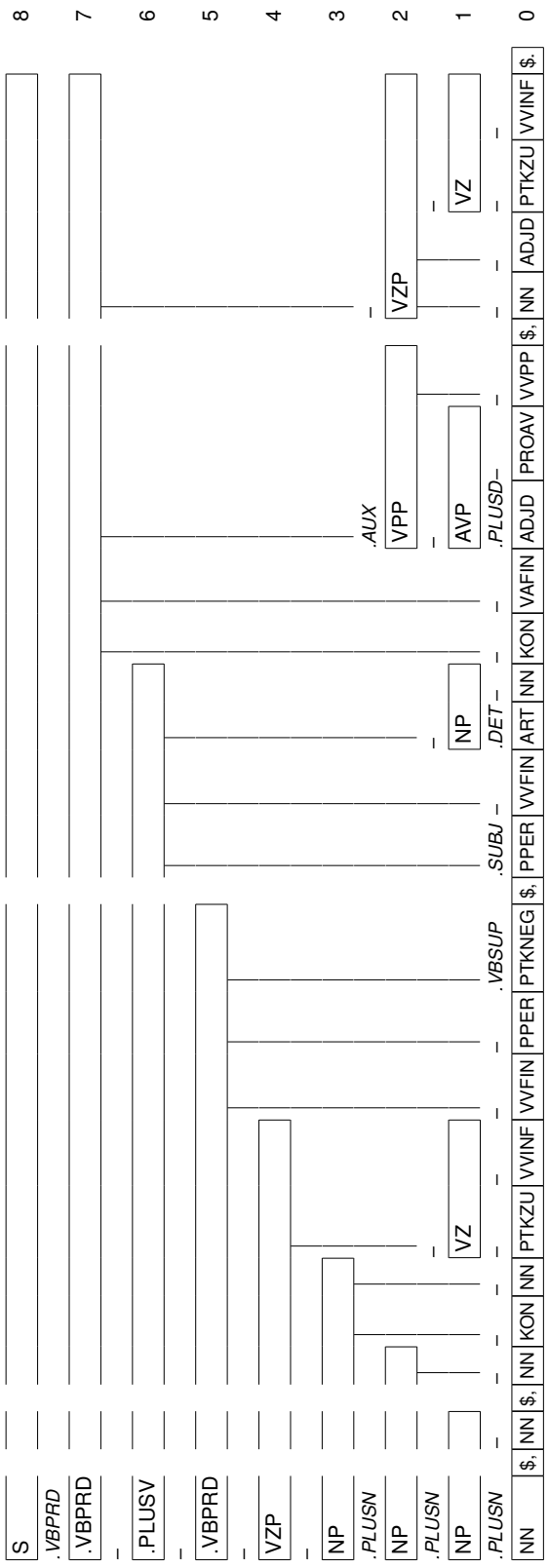


Abbildung 3.8: Analyse von Satz 3 aus NEGRA des Gojol-Parsers als Kastendiagramm. Unäre Phrasen mit Punkt-Etiketten wie .PLUSN werden als syntaktische Funktionen interpretiert.

lysiert. Vereinfacht gilt für syndetische Koordination einer Kategorie XP das Strukturschema:

$$[_{XP}\ [_{.PLUSX}\ \ldots]\ [_{KON}\ \ldots]\ [_{X}\ \ldots]\ ]$$

Für Koordinationen, welche mehr als 2 Glieder umfassen, wird .PLUSX indirekt rekursiv entsprechend dem Schema $[_{.PLUSX}\ [_{XP}\ \ldots]\ ]$ expandiert. Alle Konjunkt-Köpfe ausser dem Letzten entsprechen dem Schema $[_{.PLUSX}\ [_{X}\ \ldots]\ ]$. Dies gilt aber nicht durchgehend und diese „.PLUS"-Kategorien werden auch noch für andere iterierende Strukturen gebraucht. Eine andere Inkonsistenz liegt darin, dass es keine Kategorie für koordinierte PP gibt.

Statt KON kann auch ein Komma oder gar kein Konjunktor vorkommen. Letzteres ist meist eine grosszügige Interpretation von Koordinationskategorien oder ein Parser-Fehler. In der Dependenzstruktur selbst sind sowohl die lexikalischen Konjunktoren (im Sinne der STTS-Kategorie KON) wie auch die Kommata (wie alle Interpunktionen) isolierte Elemente, d.h. weder regierendes noch regiertes Element. Dies kann in der Abbildung 3.7 auf Seite 181 sowohl für die monosyndetische Nominalkoordination wie für die Verbkoordination abgelesen werden. Die Vorglieder, d.h. die PLUSX-Kategorien, sind immer abhängig vom Kopf des Letzglieds. Von den .PLUSX-Kategorien gibt es 4 Sorten, welche im Folgenden geordnet nach ihrer Häufigkeit kurz charakterisiert werden.

.PLUSD Dies ist weniger eine Koordinationskategorie für Determinatoren, sondern eher eine Iterationskategorie für attributive und adverbiale Adjektive, Indefinit- und Possessivpronomina in attributiver Funktion usw., was die Auflistung (189) zeigt.

(189) Expansionen der total 15058 .PLUSD-Konstituenten des Gojol-Parsers über NEGRA mit mehr als 2% Anteil (für NEGRA ergeben sich 28 Types mit einem Type-Token-Verhältnis von 1:537.8):
„ADJD" (6304, 41.9%), „ADJA" (2386, 15.8%), „CARD" (1916, 12.7%), „AVP" (1156, 7.7%), „PPOSAT" (958, 6.4%), „AJP" (956, 6.3%), „PIDAT" (436, 2.9%), „PIAT" (326, 2.2%)

.PLUSN Diese Kategorie steht für die Vorder-Konjunkte von koordinierten NP und wird im folgenden Abschnitt genau evaluiert.

(190) Expansionen der total 12218 .PLUSN-Konstituenten des Gojol-Parsers über NEGRA mit mehr als 2% Anteil (für NEGRA ergeben sich 26 Types mit einem Type-Token-Verhältnis von 1:469.9):
„NP" (6440, 52.7%), „NN" (4390, 35.9%), „.DET NN" (792, 6.5%), „TRUNC" (358, 2.9%)

.PLUSV Diese Kategorie hat 900 verschiedene Expansionen für 2730 Vorkommen, was eine enorm niedriges Type-Token-Verhältnis von 1:3.0 ergibt. Nur die Expansion „SUBJ VVFIN“ deckt im Negra-Korpus mit 56 Vorkommen mehr als 2% der Fälle ab. Bei .PLUSV handelt sich um die Vorder-Konjunkte von koordinierten finiten Sätzen (CS aus NEGRA).

.PLUSA Diese Kategorie bezeichnet nicht CAP, sondern die Erstglieder von nicht-finiten Verbalphrasen, welche von Hilfsverben (*auxiliaries*) abhängen. Gojol verwendet für seinen Parser ähnlich wie Klein und Manning (2003) eine Umbenennungsstrategie der abhängigen Verbphrase auf Grund des regierenden Hilfsverbtyps *(auxiliary split)*.

(191) Expansionen der total 502 .PLUSA-Kategorie des Gojol-Parsers über NEGRA mit mehr als 2% Anteil (für NEGRA ergeben sich 27 Types mit einem Type-Token-Verhältnis von 1:18.6):
„VVPP“ (200, 39.8%), „VVINF“ (86, 17.1%), „VIP“ (86, 17.1%), „VPP“ (62, 12.4%)

3.3.2 Evaluation der erkannten CNP des Gojol-Parsers

In diesem Abschnitt soll die Erkennungsleistung des Gojol-Parsers selektiv bezüglich der Koordinationskonstituente CNP evaluiert werden. Für eine automatische Evaluation äusserst erschwerend ist die Eigenschaft des Parsers, dass er die Token-Ebene verändert, indem er abgetrennte Verbpräfixe von ihrer ursprünglichen Stelle entfernt und sie an die mutmasslich damit zu ergänzenden Verbformen konkateniert. Diese Technik ist zwar ein nützlicher Trick, um solch häufig vorkommende diskontinuierliche Wortbestandteile in einem kontextfreien Parsingverfahren zu behandeln, allerdings sollte die zuverlässige Rekonstruierbarkeit des Inputs gegeben sein.[9] Weiter müssen zur Normalisierung alle Klammern, Anführungszeichen entfernt werden und der Parser selbst löscht alle tokeninternen Interpunktionszeichen. Für 52 Sätze des NEGRA-Korpus konnte der Parser kein Resultat ausgeben.

Für die Evaluation wurden beim Gojol-Parser die längsten Phrasen vom Typ NP genommen, welche .PLUSN-Kategorien einbetten. Beim NEGRA-Referenzkorpus sind ebenfalls die längsten CNP-Konstituenten selegiert worden, d.h. diejenigen, welche selbst keine CNP dominieren. Von den insgesamt knapp 5200 in NEGRA vorhandenen CNP bleiben damit noch gut 4800 übrig.

Evaluation der IOB-Tag Das vertikalisierte Format für die selektive Evaluation ist in Abbildung 3.9 auf der nächsten Seite gezeigt. Die Tabelle 3.9 auf der nächsten Seite zeigt, dass von den total 4846 nur leicht mehr als 1/3 korrekt identifiziert werden. Umgekehrt stimmen von den 4479 vom Gojol-Parser gefundenen CNP nur

[9] Da das Entfernen von potentiellen Verbpräfixen heuristisch geschieht, werden aufgrund der Formähnlichkeit fälschlicherweise manchmal auch der unbestimmte Artikel „ein“ oder Postpositionen wie „herum“ entfernt.

Token	NEGRA	Gojol
Sie	O	O
gehen	O	O
gewagte	B-CNP	B-CNP
Verbindungen	I-CNP	I-CNP
und	I-CNP	I-CNP
Risiken	I-CNP	I-CNP
ein	O	O
versuchen	O	O
ihre	O	O
Möglichkeiten	O	O
auszureizen	O	O

Abbildung 3.9: Selektive Projektion der CNP-Konstituenten auf die lexikalische Terminalebene für die IOB-Evaluation

Kategorie	vorhanden	gefunden	korrekt	P	R	F
CNP	4846	4479	1678	37.5	34.6	36.0

Tabelle 3.9: Evaluation der erkannten CNP des Gojol-Parsers über NEGRA. Gezählt wurden die längsten CNP. Legende: P: Precision, R: Recall, F: F-Mass

Kategorie	vorhanden	gefunden	korrekt	P	R	F
NN/NE	6666	6447	3060	47.5	45.9	46.7

Tabelle 3.10: Evaluation des Gojol-Parsers über NEGRA bezüglich der in CNP enthaltenen NN- und NE-Token. Legende: P: Precision, R: Recall, F: F-Mass

knapp 38% mit dem Referenzkorpus überein. Da hier die längsten CNP evaluiert werden, lassen sich die Zahlen bei den längeren Konstituenten nicht direkt mit den Resultaten von Chunkie in Tabelle 3.3 auf Seite 171 vergleichen.

Die Erkennungsrate von CNP in Abhängigkeit von der Anzahl dominierter lexikalischer Terminalknoten ist in der Tabelle 3.11 auf der nächsten Seite zusammengestellt. Man sieht darin, dass auch bei kurzen CNP weniger als 50% der Konstituenten gefunden werden, immerhin ist die Trefferquote bei den 3-teiligen bei knapp 66%.

Zwischen dem NEGRA-Annotationsmodell und den Resultaten des Gojol-Parsers bestehen gewisse strukturelle Unterschiede, welche bei einer Evaluation, welche exakte Phrasengrenzen verlangt, die Resultate „künstlich“ verschlechtern könnte. Um einen solchen Einfluss besser beurteilen zu können, wurde eine Evaluation gemacht, bei der das Ende der CNP mit dem Tag E-CNP markiert wurde. Die Konfusionsmatrix dieser erweiterten IOB-Chunks ist in Tabelle 3.12 auf Seite 187

CNP

Länge	P	R	F	gefunden
3	65.8	43.3	52.2	1419
5	34.8	42.3	38.2	730
4	26.5	40.1	31.9	823
6	27.0	35.8	30.8	411
7	25.2	27.7	26.4	314
10	20.0	16.0	17.8	85
8	10.6	14.2	12.2	226
9	11.5	12.1	11.8	139
13	14.3	7.0	9.4	28
11	10.9	7.2	8.7	64
12	10.3	6.7	8.1	39
16	12.5	4.8	6.9	8
14	8.3	4.3	5.7	24
18	9.1	4.3	5.9	11
2	5.2	6.5	5.7	116
15	7.7	2.8	4.1	13
17	0.0	0.0	0.0	10
19	0.0	0.0	0.0	1
20	0.0	0.0	0.0	5
21	0.0	0.0	0.0	3
22	0.0	0.0	0.0	3
23	0.0	0.0	0.0	2
24	0.0	0.0	0.0	1
25	0.0	0.0	0.0	1
26	0.0	0.0	0.0	2
33	0.0	0.0	0.0	1

Tabelle 3.11: Evaluation der vom Gojol-Parser erkannten CNP im NEGRA-Korpus aufgeschlüsselt nach der Länge. Legende: Länge: Anzahl der dominierten lexikalischen Token (keine Interpunktion), P: Precision, R: Recall, F: F-Mass, gefunden: Anzahl der vom Gojol-Parser gefundenen Phrasen

falsch	korrekt	Anzahl	Fehler in %
O	I-CNP	7831	2.60
I-CNP	O	5148	1.71
B-CNP	O	1954	0.65
O	E-CNP	1723	0.57
O	B-CNP	1667	0.55
E-CNP	O	1488	0.49
I-CNP	B-CNP	985	0.33
I-CNP	E-CNP	537	0.18
E-CNP	I-CNP	417	0.14
B-CNP	I-CNP	352	0.12
E-CNP	B-CNP	27	0.01
B-CNP	E-CNP	8	0.00

Tabelle 3.12: Konfusionsmatrix der Evaluation der IOB-Tags mit Endmarkierung für Gojol-Parser über dem NEGRA-Korpus. Beim Fehleranteil sind nur die relativen Verhältnisse relevant, da die selektive Evaluation primär O-Tags enthält, mit denen alle Nicht-CNP markiert werden.

dargestellt. Insgesamt sind es 417 Phrasen, welche zu früh enden, 537, welche zu spät enden, sowie 352 Phrasen, welche zu spät beginnen.

Evaluation bezüglich Nominalbestandteilen Ein Evaluation, welche die exakten Phrasen-Grenzen von CNP noch weniger berücksichtigt, entsteht, indem nur noch auf den NN- und NE-Token des NEGRA-Korpus ausgewertet wird. Wie in Tabelle 3.10 auf Seite 185 abgebildet, wird das F-Mass damit gut 10% besser.

3.4 Chunking- und Parsing-Ansätze von Schiehlen

3.4.1 Chunking-Ansätze

In Schiehlen (2003) werden eine ganze Reihe von Experimenten zur partiellen syntaktischen Analyse des NEGRA-Korpus vorgestellt. Dabei wird einerseits mit n-gramm-basierten Verfahren über beschrifteten Dependenzrelationen gearbeitet, welche mittels maschineller Lernverfahren auf die Aufgabe optimiert werden, andererseits verwendet Schiehlen einen Parser auf der Basis von kaskadierten endlichen Automaten, welcher mit manuell erstellten Regeln und morphologischer Information operieren. Die These von Schiehlen lautet, dass die Kombination von beiden Verfahren die optimalsten Resultate ergibt.

Um die Verfahren, welche teilweise unterspezifizierte Strukturen berechnen, vergleichen und evaluieren zu können, wählt Schiehlen ein dependenzorientiertes Repräsentationsformat. Dies definiert er ausgehend vom Evaluationsschema von Lin (1995), indem zu jedem Token *w* drei Informationen gespeichert werden:

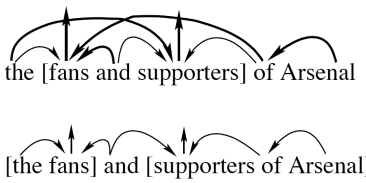


Abbildung 3.10: Verschiedene Möglichkeiten zur Dependenzkodierung nach Schiehlen (2003, 3)

1. Die Richtung, in welcher das Kopf-Token *h* liegt, von dem *w* abhängt.
2. Das Kopf-Token *h* selbst als Zeichenkette.
3. Der Abstand zum Kopf-Token in Form einer Angabe, das wievielte Kopf-Token exakt gemeint ist – typischerweise ein Abstand von 1 oder 2.

Gemäss Schiehlen funktioniert bei den n-gramm-basierten Chunking-Ansätzen das sogenannte „*nth-tag*"-Verfahren am besten, das den Kopf nicht als Token, sondern nur über seine Wortart kodiert. Alle drei der oben erwähnten Informationen können in einem Chunk-Tag notiert werden. Wenn die funktionale syntaktische Dependenz-Information wie SB aus NEGRA nicht miteinbezogen wird, ergeben sich 192 solcher Tags für das NEGRA-Korpus, mit der Dependenz-Information sind es mit 982 gut 5 Mal mehr.

Schiehlen weist ein F-Mass von 77% (Ausbeute: 75%, Präzision: 80%) nach, wenn mit dem C4.5-Entscheidungsbaumklassifikator (Quinlan 1993) mit einem Kontext von 5 Token über idealen Wortarten-Tags gechunkt wird. Unter realistischen Bedingungen mit den Wortarten eines automatischen Taggers reduziert sich das F-Mass auf 74% (Ausbeute: 72%, Präzision: 77%). Interessanterweise ist bei den n-gramm-basierten Verfahren die Präzision durchwegs höher als die Ausbeute. Über die Kohärenz der syntaktischen Struktur, welche durch diese Chunk-Tag-Klassifikation entsteht, macht Schiehlen leider keine Aussagen.

3.4.1.1 Behandlung von koordinierten Strukturen

Dafür bespricht er vertieft die Kodierung und Behandlung von koordinierten Strukturen, welche für dependenzorientiert Ansätze generell eine Herausforderung bezüglich der Bestimmung des Kopfes darstellen. Die Abbildung 3.10 zeigt im oberen Teil die Repräsentation, welche Konjunkte als gleichwertige Köpfe behandelt, wie es auch in phrasenstruktureller Repräsentation mit symmetrischer Koordination gemacht wird. Damit lässt sich die Modifikation durch den postnominalen PP-Modifikator direkt auf beiden Köpfen ausdrücken. Der Nachteil davon ist, dass die Dependenzstruktur nicht mehr projektiv ist, d.h. es entstehen „überkreuzende" Dependenzen. Die untere Repräsentation in Abbildung 3.10 steht für eine andere

Lesart, wo die Konjunkte keinen gemeinsamen Artikel und postnominalen Modifikator besitzen und jeweils lokal angebunden werden können. Um letztere Fälle strukturell von den Fällen mit gemeinsamen Modifikatoren unterscheiden zu können, verknüpft Schiehlen gemeinsame pränominale Modifikatoren mit dem letzten Konjunkt und postnominale Modifikatoren sowie Konjunktoren mit dem ersten Konjunkt. Zudem werden die Abhängigkeiten der Konjunkte mit einem „c“ markiert, z.B. SBc oder MOc. Die Ausbeute dieser mit „c“ markierten Abhängigkeiten auf die Konjunktköpfe liegt mit 13% beim „nth-tag“-Chunking sehr tief. Dies erstaunt nicht, weil die „c“-Abhängigkeiten nur knapp 0.9% aller Abhängigkeiten im Korpus ausmachen und sich zudem noch auf 22 unterschiedliche zugrunde liegende Abhängigkeiten verteilen. Ein weiterer Faktor dabei ist die durchschnittliche Token-Distanz von 7.8, welche dadurch erhöht ist, dass bei der Kodierung der Abhängigkeit beider Köpfe nach Aussen immer ein Konjunkt übersprungen werden muss.

Der kaskadierte Parser hingegen erreicht bei den „c“-Abhängigkeiten 34% bzw. 26% Ausbeute, abhängig davon, ob die unterspezifizierten Strukturen als optimal desambiguiert bzw. als durchschnittlich desambiguiert betrachet werden. Insgesamt erreicht der kaskadierende Parser über dem NEGRA-Korpus ein F-Mass von 81%, wenn unterspezifizierte Strukturen gemäss der Strategie von maximaler Anbindung aufgelöst werden und die vom Verb abhängigen Satzglieder direkt kodiert werden.

3.4.2 Parsen des NEGRA-Korpus mit der Grammatik aus der NEGRA-Baumbank nach Schiehlen

In Schiehlen (2004) werden eine Reihe von Experimenten beschrieben, welche mit dem probabilistischem CYK-basierten Parsing-System `bitpar` von Schmid (2004) durchgeführt wurden. Die Evaluationen sind sowohl mit den PARSEVAL-Metriken über die Konstituenz (im Folgenden kurz KF-Mass) als auch ähnlich wie im vorangegangenen Abschnitt mit dependenzbasierten Massen (kurz DF-Mass) vorgenommen.

Beim Evaluieren von Parsern ist der Umgang mit unbekannten Wörtern ein kleiner, aber nicht zu unterschätzender Faktor. Schiehlen testet drei Verfahren:

1. Zuordnen von Tags auf Grund von ihrer Häufigkeit im Korpus, getrennt nach Gross- und Kleinschreibung: KF-Mass: 67.2%, DF-Mass: 78.3%.

2. Zuordnen von Tags aufgrund von automatischem Tagging durch einen externen Tagger. Die Hoffnung, dass der Parser wegen den Taggingfehlern im Trainingsmaterial besser mit Taggingfehlern im Test-Set umgehen kann, wurde allerdings bezüglich Dependenz nicht erfüllt: KF-Mass 67.3%, DF-Mass: 77.7%.

3. Zuordnen von Tags auf Grund einer morphologischen Analyse, bzw. Rückgriff auf das Verfahren 1, falls ein Wort auch der morphologischen Analyse

nicht bekannt war: KF-Mass 67.4%, DF-Mass 78.1%.

Das letzte Verfahren bildet für Schiehlen die untere Limite, an der sich optimierte Varianten bewähren müssen – eine äusserst knifflige Aufgabe, wie sich herausstellte.

Experimente mit der Integration von Mutter- oder Grossmutterkategorie (*parent encoding*) im Stil von Johnson (1998) erweisen sich in Kombination mit dem NEGRA-Korpus nicht als gewinnbringend. Die Resultate verschlechtern sich insbesondere beim DF-Mass. Experimente mit der Binarisierung bzw. Ternärisierung von langen Regeln (*Markovization*), welche oft nur einmal oder zweimal auftreten und damit keine gute statistische Grundlage abgeben, verschlechtern ebenfalls, auch wenn sie in Kombination mit dem Einkodieren von dominierenden Kategorien verwendet werden. Eine deutliche Steigerung beim KF-Mass auf 72.0% (DF-Mass 74.6%) ergibt sich einzig, wenn die Mutterkategorie einkodiert wird und gleichzeitig selektiv binarisiert wird bei den Regeln auf Grund von Häufigkeitsdaten.

Um das Problem der vielen selten vorkommenden Grammatikregeln zu entschärfen oder die Regeln funktional genauer zu bestimmen, wurden noch 19 verschiedene linguistisch motivierte Verfahren ausprobiert, welche teilweise von Klein und Manning (2003) inspiriert sind. Wirklich hilfreich für beide evaluierten Masse waren aber folgende Umformulierungen der Grammatikregeln, welche vermutlich dem Problem der freien Wortstellung der Nominalphrasen am ehesten gerecht werden:

- Kasus-Information: Eine Verbesserung ergab sich, wenn Nominalphrasen und die darin befindlichen normalen Nomen[10] und Pronomen bezüglich Kasus markiert werden: KF-Mass: 71.1%, DF-Mass: 81.0.

- Subkategorisierungs-Information: Es wird ein externes Subkategorisierungslexikon für Verben, Nomen und Adjektive beigezogen.

Die Behandlung von Koordination spielt bei 2 der 19 Experimente eine Rolle. Die Ersetzung der 7 koordinierten Kategorien CXP durch XP ergab eine Verschlechterung der Resultate. Das zusätzliche Markieren von Präpositionen und Subjunktionen, welche als Konjunktor fungieren können, ergab eine minime und kaum aussagekräftige Verbesserung der Konstituentenqualität um 0.3%.

Aus den Experimenten von Schiehlen kann man die Schlussfolgerung ziehen, dass die verschiedenen Verfahren zur Präzisierung und Spezialisierung der kontextfreien Grammatikregeln mit Ausnahme der Kasus-Information bzw. Subkategorisierungsinformation grundsätzlich wenig taugen. Schiehlen selbst bleibt vorsichtig bei der Interpretation der Resultate seiner Experimente. Zurecht, wie mir scheint. Die Datenbasis der Regeln aus dem NEGRA-Korpus ist zu schmal, um generelle Folgerungen über die verwendeten Methoden zu ziehen.

[10] Über die Behandlung von engen Appositionen, welche normalerweise keine Kasusmarkierung tragen, spricht Schiehlen nicht.

Eine interessante Beobachtung zum Einfluss der verfügbaren Datenmenge bei supervisierten Lernverfahren haben Banko und Brill (2001) im Zusammenhang von Verfahren zur Wort-Desambiguierung (*confusion set desambiguation*) gemacht. Sie haben gezeigt, dass einfache Lernverfahren durch die Verwendung von grösseren Datenmengen zu Leistungen kommen können, die besser sind als das, was mit komplexen Lernverfahren aus unzureichenden Datenmengen herausgepresst werden kann. Wenn verschiedene Methoden bei 1 Million Trainingstoken ein asymptotisches Lernverhalten zeigen, bedeutet es nicht, dass bei deutlich mehr Daten diese Grenze nicht noch nach oben verschoben werden kann. Sie votieren deshalb anhand ihrer experimentellen Resultate der Wort-Desambiguierung dafür, dass ein verstärkter Effort zur Erstellung von Trainingsmaterial mindestens so lohnend ist wie die Weiterentwicklung verbesserter Lernverfahren. Entscheidend für die Mach- und Zahlbarkeit solcher Ressourcen wäre aber eine optimale automatische Unterstützung der Annotationsarbeit.

3.4.3 Evaluation des bitpar-Parsers über TIGER

3.4.3.1 Aufbereitung

Die Leistungsfähigkeit des probabilistischen CKY-Parsers `bitpar`[11] von Schmid (2004) für die Analyse von koordinierten Phrasen bezüglich Phrasenstruktur wurde in einem Evaluationsexperiment überprüft. Mit Hilfe des Graph-Umwandlungswerkzeugs der `annotate`-Software liessen sich die diskontinuierlichen Graphen in kontextfreie Strukturen mit koindizierten Spuren aufbereiten. Die syntaktischen Funktionen und Spuren sind danach gelöscht worden und in jedem Satz wurde ein TOP-Knoten ergänzt, der alle Teilstrukturen sowie die allfällig noch nicht integrierten Interpunktionen überspannt.

Das so aufbereitete TIGER-Korpus wurde in 1/10 Testmaterial (4002 Sätze) und 9/10 Trainingsmaterial (36018 Sätze) aufgeteilt, indem zufällig 1 Satz von 10 konsekutiven Sätzen ausgesondert wurde. Das Trainingsmaterial ergab 31047 verschiedene Grammatikregeln mit insgesamt 303773 Vorkommen. Da `bitpar` ein vollständiges Lexikon erwartet, wurde der `TnT`-Tagger von Brants (2000b) über dem Trainingsmaterial (ohne das Testmaterial) trainiert und danach damit das Testmaterial getaggt. Der Anteil der unbekannten Token betrug 5.9% (4247 Token). Die Verarbeitung eines Satzes benötigte auf einem Intel Duo Core Prozessor mit 2 GHz etwa 0.6 Sekunden Rechenzeit. 3 der 4002 Test-Sätze konnten nicht geparst werden.

[11] Die Software ist erhältlich unter `http://www.ims.uni-stuttgart.de/tcl/SOFTWARE/BitPar.html`.

Kategorie	vorhanden	gefunden	korrekt	Recall	Precision	F-Mass
CNP	997	991	630	63.2	63.6	63.4
CS	460	504	176	38.3	34.9	36.5
CAP	180	190	129	71.7	67.9	69.7
CVP	115	149	56	48.7	37.6	42.4
CPP	99	99	43	43.4	43.4	43.4
CO	31	27	4	12.9	14.8	13.8
CAVP	12	14	6	50.0	36.4	42.1
CAC	4	5	4	100.0	80.0	88.9
CVZ	2	2	2	100.0	100.0	100.0
CXP	1900	1981	1050	55.3	53.0	54.1
Alle	30093	30319	20713	68.8	68.3	68.6

Tabelle 3.13: Evaluation der Erkennung der Koordinationskonstituenten durch `bitpar` über TIGER. Die Zeile „CXP“ wertet alle Koordinationskonstituenten zusammen aus. Die Zeile mit der Kategorie „Alle“ evaluiert über sämtlichen syntaktischen Konstituenten ausser dem künstlich eingesetzten TOP-Knoten.

3.4.3.2 Resultate

Für die Evaluation wurde die klassische PARSEVAL-Methode[12] (Black u. a. 1991) verwendet. Diese Testverfahren ist valide, da `bitpar` als daten-basiert lernendes Verfahren exakt das Grammatikmodell von TIGER anwenden soll. Die künstlich eingefügte Konstituente TOP wurde immer ignoriert bei der Auswertung.

Die erhobenen Masse sind etikettierte Präzision (*labelled bracketing precision*), Ausbeute (*labelled bracketing recall*) und das harmonische Mittel davon, das F-Mass (*labelled bracketing f-measure*). Das Evaluationswerkzeug wurde auch selektiv jeweils für jede Koordinationskategorie separat eingesetzt, indem alle ausser die interessierende Kategorie als zu ignorieren gesetzt wurden. Die Resultate in der Tabelle 3.13 zeigt bei CNP mit dem F-Mass 63.4% deutlich bessere Resultate als der Gojol-Parser in Tabelle 3.9 auf Seite 185 mit 36.0% oder Chunkie in Tabelle 3.3 auf Seite 171 mit 53.4%. Die CAP sind mit knapp 70% F-Mass ebenfalls deutlich besser erkannt als beim Chunkie mit 51%. Die Resultate bei längeren und dementsprechend schwieriger korrekt zu berechnenden Konstituenten sind deutlich tiefer.

Auf der Ebene der Wortarten liegt die Genauigkeit bei 96.9%. Insgesamt ergibt sich für alle Konstituenten ein KF-Mass von 68.6%, was für die doppelt so grosse Trainingsmenge im Vergleich zu den 67,1% von Schiehlen (2004), der allerdings das NEGRA-Korpus benutzt hat, ein Steigerung um etwa 1.5% bedeutet.[13]

[12] Die verwendete Software `evalb` ist zur freien Verfügung unter `http://nlp.cs.nyu.edu/evalb` erhältlich.

[13] Für eine statistisch besser abgesichertere Aussage müssten allerdings für beide Korpora kreuzvalidierende Verfahren mit unterschiedlichen Trainings- und Testmengen angewendet werden, damit

Kategorie	Recall	Precision	F-Mass
CNP	68.0	67.8	67.9
CS	39.8	38.3	39.1
CAP	70.8	68.4	69.6
CVP	52.4	42.6	47.0
CPP	52.6	51.7	52.2
CO	14.3	14.3	14.3
CAVP	50.0	36.4	42.1
CAC	100.0	66.7	80.0
CVZ	100.0	100.0	100.0
CXP	58.7	56.8	57.7
Alle	72.1	71.8	71.9

Tabelle 3.14: Evaluation der Erkennung der Koordinationskonstituenten durch `bitpar` von Sätzen mit maximal 30 Token über TIGER. Die Zeile CXP wertet alle Koordinationskonstituenten zusammen aus. Die Zeile mit der Kategorie „Alle“ evaluiert über sämtlichen syntaktischen Konstituenten ausser dem künstlich eingesetzten TOP-Knoten.

Die Erkennungsleistung von Parsern wie dem `bitpar` sinkt mit zunehmender Satzlänge. Deshalb wurde zusätzlich untersucht, ob in Sätzen bis maximal 30 Token auch die Koordinationskonstituenten besser erkannt werden. Im Testkorpus haben 468 Sätze mehr als 30 Token. In der Tabelle 3.14 sieht man, dass insgesamt ein 3.3% höheres F-Mass bei den 3531 kürzeren Sätzen resultiert. Interessanterweise profitieren aber auch die häufigeren Koordinationskategorien mit Ausnahme von CAP deutlich von dieser Einschränkung, sodass das F-Mass für Koordinationskonstituenten ebenfalls um 3.6% steigt. Die Erfahrung der erheblichen Komplexität von Koordinationsstrukturen in längeren Sätzen bestätigt sich insofern quantitativ.

3.5 Hamburger Dependenzparser und Dependenzgrammatik

Im Rahmen des Projektes „papa“ (*partial parsing*) an der Universität Hamburg (Foth u. a. 2005) wurde ein umfassendes dependenzbasiertes Parsingsystem auf der Basis gewichteter Beschränkungen (*weighted constraint based dependency parsing*) entwickelt.

Im Rahmen dieses Parsing-Systems wurde auch eine umfassende Dependenzgrammatik für schriftliches Deutsch entwickelt, welche einerseits an selbst annotierten unbeschränkten Texten auf seine Abdeckung überprüft wurde. Andererseits

sich Verzerrungen mitteln lassen, welche durch zufällig speziell günstige oder ungünstige Auswahl der Trainingssätze entstehen können. Dies ist jedoch bei Schiehlen (2004) aus Gründen der Vergleichbarkeit mit anderen Evaluationen nicht der Fall.

auch durch eine Transformation des NEGRA-Korpus in dieses dependenzbasierte Grammatikmodell durch Daum u. a. (2004) in einem grösseren Rahmen empirisch evaluiert ist.

3.5.1 Hamburger Dependenz-Grammatik (HDG)

In Foth (2004b, a) ist das Grammatikmodell detailliert beschrieben. In der Abbildung 3.11 auf der nächsten Seite ist mit Satz 3 aus dem NEGRA-Korpus ein hinreichend komplexer Satz mit Koordinationsphänomenen abgebildet.

Die Grammatik verwendet beschriftete Abhängigkeiten, welche den syntaktischen Funktionen in NEGRA entsprechen. Sie zeichnet sich durch folgende Eigenschaften aus:

- Das finite Verb ist der Kopf des finiten Satzes. Nominale Subjekte (SUBJ) oder verbale Subjekte (SUBJC) sind immer davon abhängig. Die Kongruenz ist damit lokal ausgedrückt.

- Die Objekte – Akkusativobjekte (OBJA, OBJA2), Dativobjekte (OBJD), Genitivobjekte (OBJG), Objektinfinitiv, d.h. Vollverben oder meist deverbale Nomen oder Adjektive (OBJI), Präpositionalobjekt, d.h. nicht weglassbare oder präpositional nicht variierbare Präpositionalphrasen (OBJP) – hängen immer vom Vollverb ab, d.h. bei Hilfsverbkonstruktionen sind sie wie beim NEGRA-Annotationsmodell anders als das Subjekt behandelt.

- Kopf von PP sind die Präpositionen, die Abhängigkeit PN bindet den nominalen oder pronominalen Kopf (selten auch Adjektive) an.

- Konjunktionalphrasen, d.h. mit „wie“ oder „als“ eingeleitete Phrasen, werden nicht als PP aufgefasst wie etwa beim Gojol-Parser. Die abhängigen Glieder sind mit CJ verbunden, nicht mit PN. Dies wird sogar bei klassischen Konjunktoren wie „sowohl ... als auch“ so gehalten, wie die Abbildung 3.12 auf Seite 196 aus (Foth 2004a, 15) illustriert. Konsequenterweise wird deshalb auch kein koordinierendes „wie“ angesetzt.

- Begleiter, d.h. bestimmte und unbestimmte Artikel sowie attribuierende „Pronomen“, stehen in modifizierender Abhängigkeit zum Nomen (DET). Attributive Adjektive und interessanterweise auch Prädeterminatoren sind aber immer als ATTR abhängig gesetzt.

- Parenthesen und Einschübe aller Art werden zur Erhaltung der Baumstruktur pragmatisch an das letzte Wort vor der Parenthese angebunden. So werden insbesondere bei direkter oder indirekter Rede die Matrixsätze zu Parenthesen, wie in Abbildung 3.13 auf Seite 196 ersichtlich.

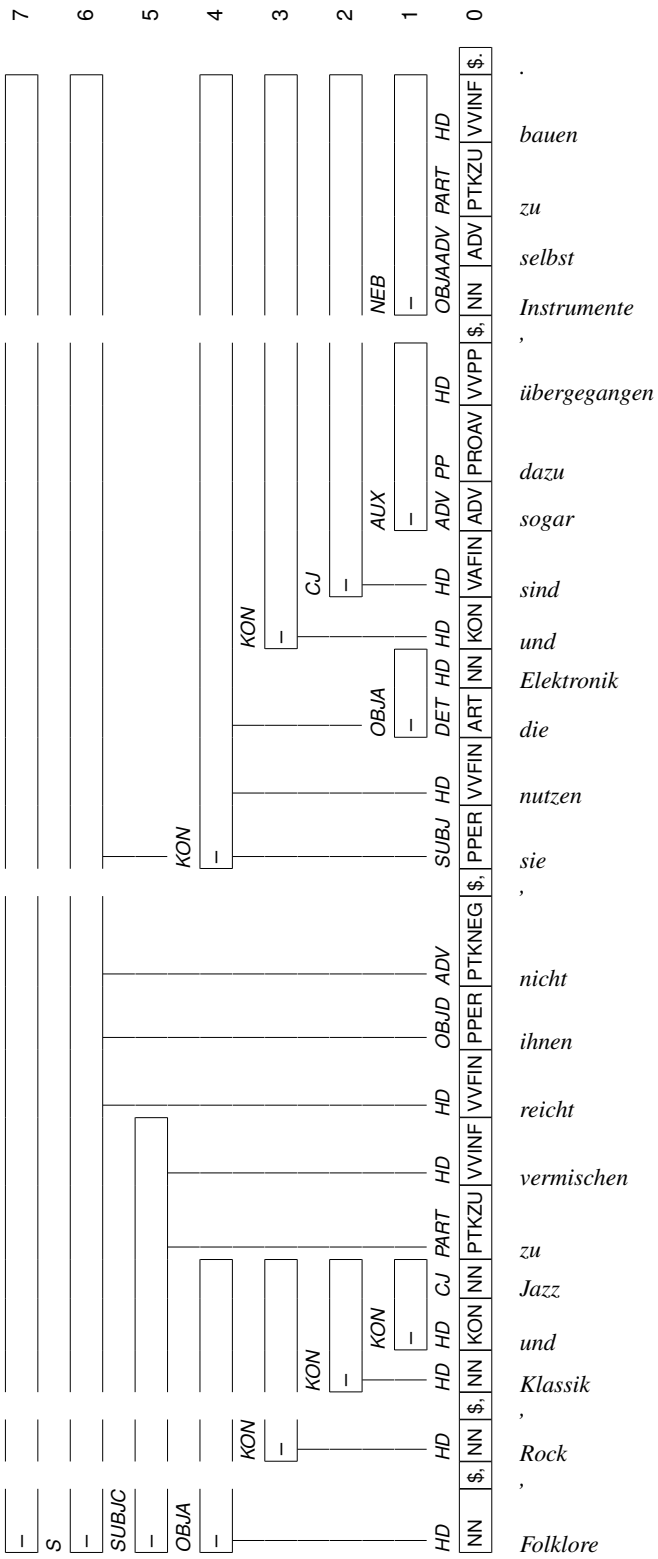


Abbildung 3.11: Dependenzrepräsentation von Satz 3 aus NEGRA konvertiert mit dem Hamburger Baumbank-Konversionswerkzeug und im Kastendiagrammformat dargestellt

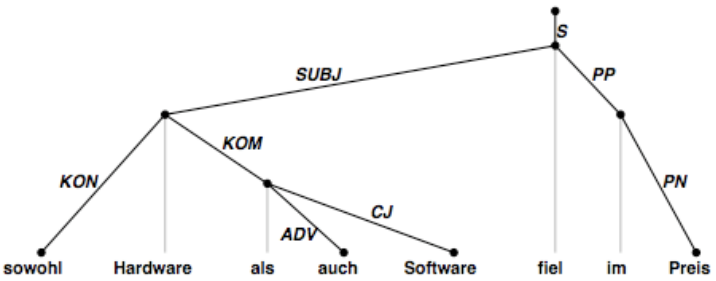


Abbildung 3.12: Die Behandlung von „sowohl ... als auch“ in der Hamburger Dependenz-Grammatik nach Foth (2004a)

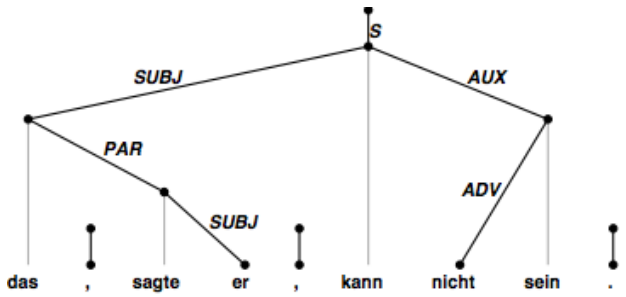


Abbildung 3.13: Die Behandlung von direkter Rede als Parenthese in der Hamburger Dependenz-Grammatik nach Foth (2004a)

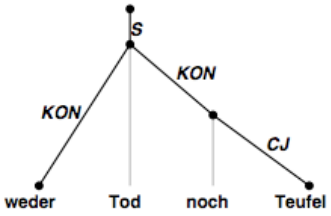


Abbildung 3.14: Die Behandlung von paarigen Konjunktoren in der Hamburger Dependenz-Grammatik nach Foth (2004a)

3.5.1.1 Koordination in der HDG

Koordinierte Strukturen werden anders als beim Gojol-Parser grundsätzlich als linksköpfig betrachtet[14]. Bei syndetischen Koordinationen hängt der Konjunktor vom Erstglied mit dem Label KON ab. Da die Interpunktion in der Grammatik nicht in die Struktur eingebaut ist, steht koordinierendes Komma bei monosyndetischen oder asyndetischen Koordinationen nicht als Verbindungsglied für KON zur Verfügung. Deshalb werden bei Koordinationsbestandteilen, welche nicht durch lexikalische Konjunktoren verknüpft sind, die Konjunkte selbst mit einer KON-Kante verknüpft Das Letztglied in monosyndetischen Koordinationen wird jedoch anders als alle vorangehenden Konjunkte mit dem Kürzel CJ angebunden. Der lexikalische Status des letzten Konjunkts ist dabei unerheblich.

Bei paarigen Konjunktoren wird wie in Abbildung 3.14 ersichtlich der einleitende Konjunktor mit KON vom ersten Konjunkt abhängig gemacht.

Der Aufbau der koordinierten Strukturen ist in der HDG-Grammatik stark asymmetrisch. Die Gleichschaltung der syntaktischen Dependenz von Konjunktoren und Konjunkten, welche nicht als Letztglied fungieren, stellt keine transparente strukturelle Kodierung dar. Letztlich lassen sich aber die Konjunktoren immer identifizieren, weil sie eine geschlossene lexikalische Klasse darstellen. Man kann sich fragen, warum es das Label CJ braucht, wenn KON dieselbe Funktion bei asyndetischen Koordinationen wahrnimmt.

Als Beispiel für die Resultate des Papa-Parsers[15] illustriert Abbildung 3.15 auf der nächsten Seite die textuelle Ausgabe. Die Stemmarepräsentation in der Tradition der Dependenzgrammatik findet sich dazu in Abbildung 3.16 auf der nächsten Seite.

3.5.2 Evaluation des WCDG-Parsers

In Daum u. a. (2003) wird die Parsing-Leistung auf der Basis der korrekt zugewiesenen Dependenz-Tags auf einem Zeitungstext-Korpus aus 1845 Sätzen (durch-

[14] Weitere Ansätze zur dependenzgrammatischen Behandlung von Koordinationen sind in Lobin (1993, 84ff.) besprochen.

[15] Das System kann unter `http://nats-www.informatik.uni-hamburg.de/parse/TWiki/CdgParserDemo` ausprobiert und ist frei verfügbar.

```
Edges:
   000        SYN: das_ART_nom(0-1)--DET-->Kind_NN(1-2)
   002        SYN: Kind_NN(1-2)--SUBJ-->gingen_VVFIN_third_past(8-9)
   004        SYN: ,_$,(2-3)---->NIL
   006        SYN: der_ART_sg_nom(3-4)--DET-->Hund_NN(4-5)
   008        SYN: Hund_NN(4-5)--KON-->Kind_NN(1-2)
   010        SYN: und_KON(5-6)--KON-->Hund_NN(4-5)
   012        SYN: die_ART_sg_nom(6-7)--DET-->Frau_NN(7-8)
   014        SYN: Frau_NN(7-8)--CJ-->und_KON(5-6)
   016        SYN: gingen_VVFIN_third_past(8-9)--S-->NIL
   018        SYN: zusammen_ADV(9-10)--ADV-->spazieren_VVINF(10-11)
   020        SYN: spazieren_VVINF(10-11)--OBJI-->gingen_VVFIN_third_past(8-9)
   022        SYN: ._$.(11-12)---->NIL

Violations:
      008 : 9.709e-01 : mod-Distanz
      002 : 9.930e-01 : Komplementdistanz
      008 : 9.950e-01 : mod
      010 : 9.950e-01 : mod
      018 : 9.950e-01 : mod
      014 : 9.980e-01 : Komplementdistanz
      020 : 9.980e-01 : Komplementdistanz
      016 : 9.990e-01 : pl
      018 : 9.999e-01 : direction
```

Abbildung 3.15: Textuelle Ausgabe des Papa-Parsers

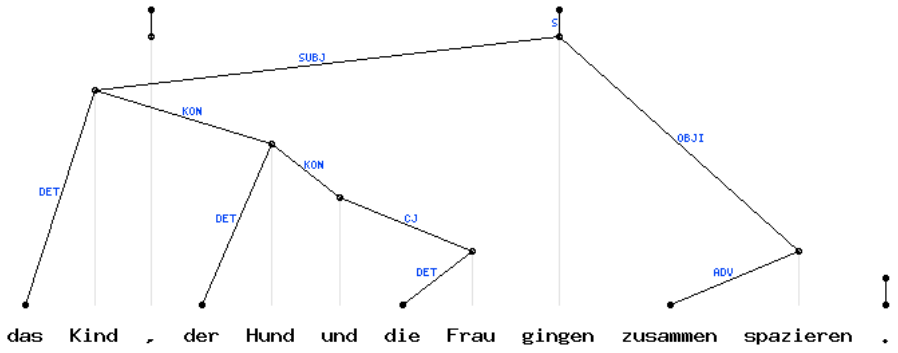


Abbildung 3.16: Stemmarepräsentation der Resultate des CDG-Parsers

schnittlich 24 Token pro Satz) erhoben. Die Leistung des reinen Constraint-Systems wird mit 50.7% Recall angegeben. Wenn die Information von einem Wortarten-Tagger und einem einfachen Chunker mit den geeigneten Gewichten einbezogen wird, erhöht sich der Recall auf 75.7% bei gleichzeitiger Reduktion des in diesem System sehr hohen Berechnungsaufwandes[16] um mehr als die Hälfte.

In Foth u. a. (2004) wird die Leistungsfähigkeit des Dependenz-Parsers gemessen an knapp 1000 Sätzen von NEGRA, welche automatisch ins Dependenzgrammatik-Format umgewandelt wurden. Dabei wird der Aufbau der Dependenzstruktur anders als in Daum u. a. (2003, 100) in 3 Phasen inkrementell berechnet. Für die korrekt etikettierten Dependenztags wird dabei ein Recall von 87% erreicht. Leider sind keine kategorienspezifische Evaluationsresultate publiziert, welche Aussagen bezüglich der Erkennung von koordinierten Strukturen zulassen.

[16] Der in Daum u. a. (2003, 100) angegebene Schnitt von 134 Sekunden Rechenzeit pro Satz bestätigte sich auch in eigenen Experimenten mit dem System auf Rechnern wie in der Fussnote auf Seite 170 angegeben; die Satzlänge spielt für die Komplexität des Problems die entscheidende Rolle, da die Berechnungszeit exponentiell davon abhängt (Foth u. a. 2005).

Kapitel 4

Desambiguierung koordinierter Strukturen für partielle syntaktische Analyse

In diesem Kapitel werden Eigenschaften von koordinierten Strukturen getestet, welche für die Desambiguierung von koordinierten Strukturen in partiellem Parsing potentiell nützlich sein können.

4.1 Effekte der Kopfdistanz

Wie weit auseinander liegen die Köpfe von benachbarten Konjunkten? Um diese Frage anhand der verwendeten Baumbanken beantworten zu können, müssen zunächst die Köpfe in allen koordinierten Konjunkten bestimmt werden.

4.1.1 Kopfdistanz in CNP

Da in NEGRA und TIGER für NP keine explizite Annotation von Köpfen mit einem Funktionslabel HD gemacht sind, muss dafür eine Heuristik angewendet werden.

4.1.1.1 Heuristiken zur Kopfbestimmung in NP

Die Bestimmung von Köpfen ist für NEGRA im Rahmen von Baumbank-Konversionen in den LTAG-Grammatikformalismus bei Frank (2001) beschrieben. Die in Bohnet (2003) präsentierte partielle Umwandlung von NEGRA ins Dependenzformat benutzt das generische graphbasierte Grammatikentwicklungswerkzeug MATE für die Kopfbestimmung. Diese Transformation lässt sich über dem kanonischen PROLOG-Baumformat (vgl. Kapitel 6.2 auf Seite 265) gut durchführen. Wie schon in Abschnitt 2.4.2 auf Seite 63 diskutiert, ist die syntaktische Funktion

NK (Nominalkern) mehrdeutig und die NK stehen innerhalb einer NP mit wenigen Ausnahmen direkt nebeneinander.

Nominalphrasen in der GDS-Grammatik Im pränominalen Bereich setzt Zifonun u. a. (1997, 2069) die folgenden optionalen Komponenten an:

- Determinative (D): Artikel, Possessivbegleiter etc. Sogenannte Prädeterminatoren wie die unflektierten „manch“ oder „all“ in „manch ein Haus“ oder „all die Menschen“ werden ebenfalls dazugezählt.
- Genitiv-Attribute (G)
- Adjektiv-Attribute (A)
- Erweiterungsnomina[1] (E): Als nomen varians „Herrn“ in „wegen Herrn Grüningers Erfolg“ oder nomen invarians „Paul“ in „wegen Paul Grüningers Erfolg“.
- Präpositional-Attribute (P): Eher selten und typischerweise in Funktion als Herkunftsausdrücke wie „von Chomsky das Buch“ oder „aus Panama einen Hut“.

Für diese Elemente wird in der GDS-Grammatik ein Set von Einschränkungen formuliert, was die Kombinierbarkeit, Abfolge und Iterierbarkeit betrifft. Unverträglich sind insbesondere Präpositional- und Genitiv-Attribute. In regulärer Kurznotation [2] ausgedrückt sehen die erlaubten Kombinationen dann so aus:

(192) a. `P ? (((D? D)? A*)| E?) N`
b. `(((D? D | G)? A*)| E?) N`

Die postnominalen Komponenten umfassen nach GDS:

- Erweiterungsnomina (E)
- Genitiv-Attribute (G)
- Präpositionalphrasen oder Konjunktionalphrasen mit „als“ (P)
- Adverbien (Adv): typischerweise temporal oder lokal
- finite und nicht-finite Nebensätze (NS)
- weite Appositionen (APP).

Die Einschränkungen bezüglich Kombinierbarkeit und Reihenfolge ergeben folgendes Muster:

(193) `(G | E)? PP* Adv? NS? APP?`

[1] Pränominale Erweiterungsnomina (in andern Grammatiken wie Helbig und Buscha (1991) als 'enge Apposition' bezeichnet), welche vorwiegend aus Vornamen, Verwandschafts- und Berufsbezeichnungen, Titeln und Anreden bestehen, sind iterierbar und auch koordinierbar.

[2] |: Alternativen; *: beliebige Wiederholung; ?: Optionalität

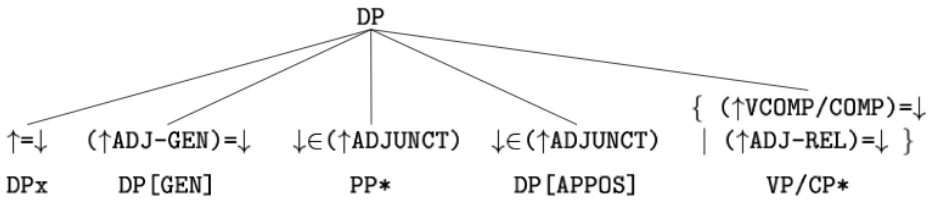


Abbildung 4.1: Postnominale Komponenten in der LFG-Analyse nach (Dipper 2003, 283). Ihre ausführliche formale Grammatik mit den exakten Spezifikationen findet sich im Anhang D von Dipper (2003).

Postnominale Struktur in TIGER und NEGRA Die Tabelle 4.1 auf der nächsten Seite zeigt die Verteilung der postnominalen Tochterkonstituenten gezählt ab der letzten NK-Konstituente. Die Bedeutung der Funktionsbezeichnungen ist im Anhang A.3 auf Seite 300 verzeichnet. Enge Appositionen sind nicht miteinbezogen, da sie ebenfalls als NK annotiert werden. Relativ häufig tauchen Appositionen (APP) vor den postnominalen Modifikatoren (MNR) oder Relativsätzen (RC) auf, was oft durch typisch schriftliche Informationsergänzungen in Klammern bedingt ist.

Nominalphrasen in der LFG-Grammatik von Dipper Eine ausführliche Implementation der deutschen Nominalphrase als DP-Analyse in der Grammatikentwicklungsumgebung XLE (Butt u. a. 1999) im Rahmen eines LFG-Ansatzes gibt Dipper (2003). Ihre postnominale Reihenfolge sieht etwas anders aus als in der GDS-Grammatik, wie in Abbildung 4.1 ersichtlich ist. Tatsächlich finden sich Appositionen in NEGRA und TIGER ja auch vor Relativsätzen. Aus Effizienzgründen beschränkt Dipper postnominale PP in ihrer Grammatik auf maximal 2 Wiederholungen (die Notation PP* in Abbildung 4.1 ist in diesem Punkt ungenau). Wenn man die Tabelle 4.1 betrachtet, ist das eine vernünftige Einschränkung.

Mehrfache nominale NK-Füllung Die Tabelle 4.2 auf Seite 206 zeigt die Verteilung aller Kernbestandteile, welche aus normalen Substantiven (NN), ein- und mehrteiligen Eigennamen (NE, MPN, PN), Massangaben (NM), koordinierten und normalen Nominalphrasen (CNP, NP) sowie nicht-initialen Kardinalzahlen (CARD) in NK-Funktion bestehen. Das häufigste und eindeutige Kombinationsmuster ist Substantiv gefolgt von Eigenname, wobei letzterer als enge Apposition (nomen invarians in GDS-Terminologie) funktioniert. Zwei normale Substantive nacheinander sind mehrdeutig, wenn keine morphologische Information vorhanden ist.

Nominalisierte Kardinalzahlen wie in (194a) oder Massangaben wie in (194b) sind typisch. Adverbiale wie in (194c), wo eine unflektierte Zeitangabe auf Substantive wie „Anfang“, „Ende“ oder „Mitte“ folgt, werden von Bohnet (2003) als linksköpfig betrachtet.

NEGRA				TIGER			
in %	Anzahl	Funktion	k.	in %	Anzahl	Funktion	k.
64.3	27177	–	64	63.5	55155	–	64
13.0	5501	MNR	77	12.1	10552	MNR	76
8.6	3628	GR	86	9.5	8225	AG	85
3.3	1385	RC	89	3.5	3002	RC	89
3.1	1290	APP	92	1.8	1531	APP	90
1.3	546	PG	94	1.6	1351	PG	92
1.0	418	PH RE	95	1.0	857	OC	93
0.9	363	OC	96	0.9	745	AG MNR	94
0.8	330	GR MNR	96	0.8	736	OP	95
0.7	276	MNR MNR	97	0.8	695	PH RE	96
0.4	185	MO	97	0.7	641	PAR	96
0.4	179	GR APP	98	0.5	472	AG APP	97
0.3	130	MNR RC	98	0.5	419	MNR MNR	97
0.3	119	CC	98	0.3	288	MO	98
0.3	114	MNR APP	99	0.3	278	CC	98
0.2	87	RE PH	99	0.3	256	MNR RC	98
0.2	76	GR RC	99	0.2	188	AG RC	98
0.2	64	PG MNR	99	0.2	148	MNR APP	98
0.1	41	GR OC	99	0.2	144	PG MNR	99
0.1	37	GR MNR MNR	100	0.1	92	AG OC	99
0.1	31	APP RC	100	0.1	79	RE PH	99
0.1	28	APP MNR	100	0.1	79	AG OP	99
0.1	22	APP APP	100	0.1	74	CVC	99
0.0	20	PG RC	100	0.1	57	PG APP	99
0.0	17	RC APP	100	0.1	55	MNR PAR	99
0.0	15	PG APP	100	0.1	47	PG RC	99
				0.1	47	AG MNR MNR	99
				0.1	46	APP RC	100
				0.1	45	AG PAR	100
				0.0	38	APP MNR	100
				0.0	34	RC APP	100
				0.0	23	MNR OP	100
				0.0	19	OP MNR	100
				0.0	16	PG OC	100
				0.0	16	MNR MO	100
				0.0	16	MNR MNR MNR	100
				0.0	15	PG OP	100
				0.0	15	MO MNR	100
				0.0	15	MNR OC	100

Tabelle 4.1: Verteilung der postnominalen Funktionen hinter NK-Elementen von NP in NEGRA und TIGER. Erhoben wurde ab der letzten Tochter, welche in NK-Funktion steht. Der Eintrag „–“ bedeutet, dass keine postnominalen Elemente nachfolgen. Gezeigt werden Mindestvorkommen ab 15, die kumulative Spalte bezieht sich aber auf alle Vorkommen. Lesebeispiel: In 5501 Fällen kommt in NEGRA innerhalb von einer NP anschliessend an die letzte mit dem Funktionslabel NK beschriftete Tochterkonstituente eine Schwester in MNR-Funktion vor.

(194) a. [$_{NP}$ [$_{NN-NK}$ Zehntausende] [$_{NN-NK}$ Demonstranten]] trugen die Bahren [. . .] [T$_{360}$]

b. [$_{NP}$ [$_{AP}$ Etwa 700 000] [$_{NN-NK}$ Quadratmeter] [$_{NN-NK}$ Büros]] sind nach Schätzung der Bank derzeit in Bau; [T$_{302}$]

c. Industrieminister Andrianopoulos verschickte deshalb [$_{NP}$ [$_{NN-NK}$ Anfang] [$_{NN-NK}$ Januar] [$_{AG}$ dieses Jahres]] ein Rundschreiben [. . .] [T$_{784}$]

Kardinalzahlen vor nominalen Füllungen sind meist Zahladjektive und werden deshalb nicht angezeigt in der Tabelle 4.2 auf der nächsten Seite. Wenn sie nach nominalen Bestandteilen stehen, sind es meist enge Appositionen, ebenso wie nachfolgende CNP.

Kopfheuristik für NP und eingebettete in PP flach eingebettete NP Das implementierte Verfahren teilt die Konstituenten auf Grund ihrer Funktion und ihrer kategorialer Füllung in zwei sich überschneidende Klassen ein.

1. Nominale Tags und Phrasen mit potentieller Kernfunktion: NN, NE, NNE, CARD, PPER, PIS, PDS, PRELS, PPOSS, XY, FM; NP, CNP, MPN, PN

2. Postnominale Tags und Phrasen mit potentieller Appositionsfunktion: NE, NNE, CARD, XY, FM, MPN, PN sowie CNP

Die folgenden Regeln für die Zuordnung der Kopffunktion werden von links nach rechts in den Tochterkonstituenten angewendet:

1. Falls nur ein Element NK-Funktion aufweist, gilt es als Kopf.[3]

2. Falls zwischen 2 benachbarten nominalen Konstituenten in NK-Funktion ein Tokendistanz > 1 besteht, gilt die 1. Konstituente als Kopf.

3. Falls die NP aus genau 2 nominalen NK-Konstituenten besteht, gilt die 1. Konstituente als Kopf.

4. Sonst gilt diejenige nominale, aber nicht potentiell appositive NK-Konstituente als Kopf, welche am weitesten rechts steht.

Bei mehrteiligen Eigennamen zählen alle Token zusammen als komplexer Kopf.

4.1.1.2 Paarweise Kopfdistanz

Die Tabelle 4.3 auf Seite 207 zeigt, wie weit auseinander in Token gemessen, adjazente Köpfe innerhalb einer CNP auftreten. Bei mehrteiligen Eigennamen wurde beim Kopf der linksstehenden Konstituente von deren Token, das am weitesten rechts steht, bis zum Token, das am weitesten links steht, von der rechten Konstituente gemessen. Die Bandbreite geht von einer lexikalischen Distanz von 1 bis

[3] Komplexe enge Annotationen wie Titel von Filmen und andere Benennungen werden oft durch Anführungszeichen oder Klammern abgetrennt.

NEGRA				TIGER			
in %	Anzahl	Konstit.	kum.	in %	Anzahl	Konstit.	kum.
84.2	34115	N	84	84.2	70464	N	84
4.8	1965	N E	89	6.1	5143	E	90
4.6	1879	E	94	4.9	4084	N E	95
2.3	948	CNP	96	2.0	1650	CNP	97
0.9	377	N N	97	0.7	606	N N	98
0.8	328	N C	98	0.5	400	NM N	98
0.4	163	NM N	98	0.4	374	N C	99
0.4	155	N NP	98	0.2	202	N CNP	99
0.4	149	N CNP	99	0.2	165	N NP	99
0.2	78	E E	99	0.1	106	E N	99
0.1	49	E C	99	0.1	76	CNP E	99
0.1	45	E N	99	0.1	70	NM	100
0.1	44	N NM	99	0.1	51	E C	100
0.1	32	CNP E	99	0.1	44	N NM	100
0.1	29	N N E	100	0.0	34	NP N	100
0.1	22	NP N	100	0.0	31	NP	100
0.0	20	NM	100	0.0	30	NM N N	100
0.0	18	N C E	100	0.0	22	N N E	100
0.0	18	NP	100	0.0	12	CNP N	100
0.0	17	N E E	100	0.0	9	E E	100
0.0	16	NM N N	100	0.0	7	NM CNP	100
0.0	10	CNP N	100	0.0	6	N N C	100
0.0	8	N N C	100	0.0	6	N C N	100
0.0	7	N E C	100	0.0	6	E N E	100
0.0	6	N N N	100	0.0	6	CNP NP	100
0.0	6	E N E	100	0.0	5	E C E	100
0.0	6	E E E	100	0.0	4	E N N	100
0.0	4	E N N	100				

Tabelle 4.2: Verteilung der nominalen kategorialen Füllung der NK-Konstituenten von NP in NEGRA und TIGER. Gezeigt werden Mindestvorkommen von 4, die kumulative Spalte bezieht sich aber auf alle Vorkommen. Aus Platzgründen und zur Vereinheitlichung wurden die NEGRA-Etiketten abgekürzt: C: CARD, N: NN, E: NE, MPN, PN. Legende: Konstit. = Konstituentenfüllung.
Lesebeispiel: In 1965 Fällen kommt in NEGRA innerhalb von einer NP ein normales Substantiv (N) gefolgt von einem Eigennamen (E) vor, welche beide die Funktion NK tragen.

NEGRA (total 6056 Paare) (Mittelwert: 2.72 ±2.12)				TIGER (total 11032 Paare) (Mittelwert: 2.8 ±2.24)			
in %	Anzahl	Distanz	kum.	in %	Anzahl	Distanz	kum.
48.6	2942	2	49	50.5	5568	2	50
17.3	1050	1	66	15.3	1691	3	66
15.4	930	3	81	15.3	1683	1	81
7.3	443	4	89	6.6	733	4	88
3.9	236	5	92	3.9	426	5	92
2.7	161	6	95	2.7	297	6	94
1.5	91	7	97	1.7	190	7	96
1.1	64	8	98	1.0	111	8	97
0.6	37	9	98	1.0	110	9	98
0.4	27	10	99	0.5	50	11	98
0.3	18	13	99	0.4	46	10	99
0.3	17	11	99	0.4	39	12	99
0.2	11	12	100	0.1	16	13	99
0.1	7	14	100	0.1	14	14	100
				0.1	13	15	100
				0.1	8	17	100
				0.1	7	16	100

Tabelle 4.3: Verteilung der paarweisen lexikalischen Token-Distanz zwischen den Köpfen von CNP in NEGRA und TIGER. Gezeigt werden die Fälle mit mindestens 5 Vorkommen.

zu 38 bei NEGRA und von 1 bis zu 39 bei TIGER. Ein Abstand von 2 stellt den Normalfall dar. Die abnehmende Häufigkeit korreliert mit der Distanz. Die Kumulation der Werte zeigt, dass innerhalb einer Distanz von 7 Token knapp 95% aller Fälle abgedeckt sind.

4.1.1.3 Konjunktoranzahl zwischen Köpfen von CNP

Ein Merkmal, welches auf der getaggten Tokenebene erhoben werden kann, ist die Anzahl der Kommas und lexikalischen Konjunktoren (KON), welche zwischen 2 Köpfen erscheinen. Da vor einigen Konjunktoren Komma stehen muss, wurden nur diejenigen Kommas gezählt, denen nicht unmittelbar ein lexikalischer Konjunktor nachfolgt. Die Tabellenzusammenstellung 4.4 auf der nächsten Seite zeigt die Verhältnisse für NEGRA und TIGER in Bezug auf die reine Tokendistanz. Bis und mit Distanz 9 sind maximal 1 Koordinationsmittel zwischen den Konjunkte in der Mehrheit.

Die Distanz 0, d.h. weder Komma noch Konjunktor als Verknüpfungsmittel, stellt einen Sonderfall dar und ergibt sich aus Tokenisierungsfehlern, der Annotation von abkürzenden Reihungsausdrücken wie „usw.“ als Konjunkt sowie bei

NEGRA

Konj.	Anzahl	in %	Dist.	Anzahl	in %	kum.
1	5732	94.6	2	3581	59.1	59
			3	1001	16.5	76
			5-9	633	10.5	86
			4	455	7.5	94
			10-	62	1.0	95
2	155	2.6	5-9	84	1.4	96
			10-	64	1.1	97
			4	7	0.1	97
0	124	2.0	2	92	1.5	99
			3	14	0.2	99
			1	9	0.1	99
			4	6	0.1	99
			5-9	3	0.0	99
3	32	0.5	10-	25	0.4	100
			5-9	7	0.1	100
4-	13	0.2	10-	13	0.2	100

TIGER

Konj.	Anzahl	in %	Dist.	Anzahl	in %	kum.
1	10470	94.9	2	6565	59.5	60
			3	1851	16.8	76
			5-9	1162	10.5	87
			4	769	7.0	94
			10-	123	1.1	95
2	256	2.3	5-9	141	1.3	96
			10-	111	1.0	97
			4	3	0.0	97
			3	1	0.0	97
0	212	1.9	2	133	1.2	98
			4	32	0.3	99
			1	25	0.2	99
			3	13	0.1	99
			5-9	6	0.1	99
			10-	3	0.0	99
3	59	0.5	10-	49	0.4	100
			5-9	10	0.1	100
4-	35	0.3	10-	35	0.3	100

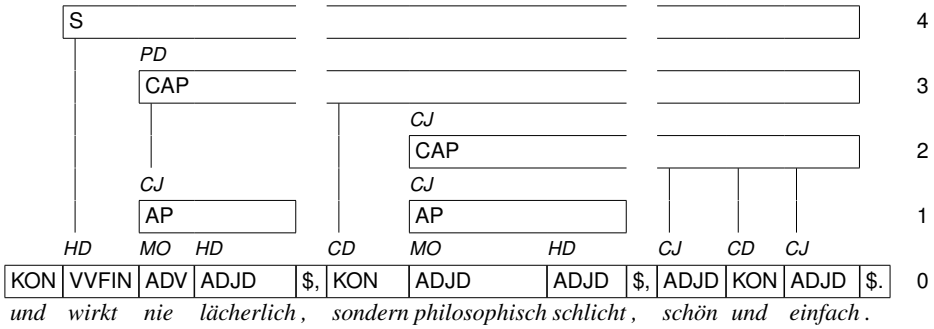
Tabelle 4.4: Verhältnis der Konjunktoranzahl (Kommas und lexikalische Konjunktore) zur reinen Tokendistanz zwischen Köpfen von CNP in NEGRA und TIGER. Legende: Dist. = Tokendistanz(bereiche) inklusive Interpunktion; Konj. = Summe aller Kommas sowie der Token mit Wortart KON, welche zwischen den Köpfen erscheinen (nicht gezählt wurden Kommas, wenn sie direkt einem Konjunktor vorangehen). Lesebeispiel: Zwischen 36 Paaren von Köpfen aus benachbarten Konjunkttöchtern, welche 5-9 Token voneinander entfernt sind, ist nur ein Konjunktor vorhanden auf der Tokenebene.

Namensbezeichnungen, welche als kommalose Reihung gebaut sind, wie etwa in „das Institut Jugend Film Fernsehen“.

4.1.2 Kopfdistanz in CAP

Die Köpfe in den Konjunkten von CAP sind entweder terminale Tochterknoten, welche direkt mit der Funktion CJ dominiert werden, oder AP, deren Köpfe explizit mit der Funktion HD markiert sind. Falls das Konjunkt selbst eine CAP ist wie in (195), stellt sich die Frage, ob für die paarweise Auswertung die Konjunkte aus diesen komplex geschachtelten Koordinationen als zusammengehörig interpretiert werden sollen. Da hier die Bezüge zwischen den Köpfen interessieren, sind im Folgenden die Köpfe von direkt rekursiv verschachtelten CAP miteinander verrechnet[4]. Im Beispiel (195) stehen somit die Kopfpaare „lächerlich/schlicht“, „schlicht/schön“ sowie „schön/einfach“ in untersuchten paarweisen Verhältnissen zueinander.

(195) Und doch bezaubert ihre Naivität



[N_{24}]

Da in NEGRA innerhalb einer CAP kein mehrteiliges Adjektive (MTA) und in TIGER nur eines vorkommt, lohnt sich eine Unterscheidung wie bei den mehrteiligen Eigennamen nicht. Der Vollständigkeit halber sei festgehalten, dass bei den Distanzberechnungen nach links von dem am weitesten links stehenden Token eines MTA gezählt wurde und umgekehrt in Richtung rechts von dem am weitesten rechts stehenden Token.

Wie verteilen sich die Kopfpaare auf die Wortarten? Die Auflistung der Paare aus NEGRA in (196) zeigt etwas andere Anteile als die Paare aus TIGER in (197). Wie in Abschnitt 2.2.1.3 auf Seite 20 diskutiert, sind bei NEGRA die pränominalen asyndetischen Koordinationen weniger konsequent annotiert. Durch das Weglassen der Texte aus den Veranstaltungshinweise weist TIGER weniger Konstruktionen mit CARD auf.

(196) Verteilung der Wortarten der Kopfpaare in NEGRA (mindestens 4 Vorkommen):

[4] Solche Annotation sind allerdings selten. In NEGRA gibt es 3 und in TIGER 4 Fälle.

„ADJA ADJA“ (407, 42.5%), „CARD CARD“ (255, 26.6%), „ADJD ADJD“ (250, 26.1%), „TRUNC ADJA“ (23, 2.4%), „TRUNC ADJD“ (7, 0.7%)

(197) Verteilung der Wortarten der Kopfpaare in TIGER (mindestens 4 Vorkommen):
„ADJA ADJA“ (1169, 58.3%), „ADJD ADJD“ (479, 23.9%), „CARD CARD“ (244, 12.2%), „TRUNC ADJA“ (59, 2.9%), „TRUNC ADJD“ (16, 0.8%), „PIAT ADJA“ (9, 0.4%), „PIAT PIAT“ (6, 0.3%), „ART ADJA“ (5, 0.2%), „ART CARD“ (4, 0.2%)

Die Tabellenzusammenstellung 4.5 auf der nächsten Seite zeigt die Verteilung der paarweisen lexikalischen Kopfdistanz in NEGRA und TIGER insgesamt und aufgeschlüsselt für reine Paare aus ADJA und ADJD. Die Distanz von 1 bedeutet dabei, dass die beiden Köpfe direkt nebeneinander liegen oder nur durch Interpunktionstoken voneinander getrennt sind. Im Beispiel (195) haben „lächerlich“ und „schlicht“ eine Distanz von 3, „schlicht“ und „schön“ eine Distanz von 1 und „schön“ und „einfach“ eine Distanz von 2. Über 95% aller Fälle liegen bei NEGRA innerhalb einer Distanz von 4, bei TIGER innerhalb der Distanz 5. Attributive Adjektive stehen im Mittel etwas näher beieinander. Bei den ADJD ist asyndetische Koordination, d.h. Distanz von 1, deutlich seltener.

4.1.2.1 Konjunktoranzahl zwischen Köpfen von CAP

Die Anzahl der Kommas und lexikalischen Konjunktoren (KON), welche zwischen 2 Köpfen erscheinen, ist in der Tabellenzusammenstellung 4.6 auf Seite 212 für NEGRA und TIGER in Bezug zur reinen Token-Distanz erhoben. Da vor einigen Konjunktoren Komma stehen muss, wurden nur diejenigen Kommas gezählt, denen nicht unmittelbar ein lexikalischen Konjunktor nachfolgt. Es zeigt sich, dass sogar bei grösseren Distanzen bis 9 Token zwischen den Köpfen selten mehr als ein Komma oder Konjunktor dazwischen auftaucht. Erst ab dem Abstand 10 konkurrenzieren intermittierende Kommas und Konjunktoren, welche nichts mit den CAP-Konjunktoren zu tun haben.

4.1.3 Kopfdistanz in CPP

In der Tabelle 4.7 auf Seite 213 ist die Verteilung der lexikalischen Distanz zwischen den Haupt-Präpositionen von koordinierten CPP dargestellt. Als Hauptpräposition betrachtet wird das am weitesten links stehende Token, das die Funktion AC aufweist. Die Distanzen sind im Schnitt fast doppelt so hoch wie bei den CNP, was sich nicht allein durch die zusätzlichen Präpositionen oder Postpositionen erklären lässt. Die Komplexität dieser Koordinationskategorie hat sich auch in Abschnitt 3.4.3 auf Seite 191 in der Evaluation der `bitpar`-Resultate gezeigt, wo CPP schlechte Erkennungsraten gezeigt haben. Die Werte sind im Vergleich zu den

NEGRA

Alle Paare (total 957 Paare)
(Mittelwert: 2.16 ±1.2)

in %	Anzahl	Distanz	kum.
76.1	728	2	76
12.1	116	1	88
6.2	59	3	94
1.9	18	4	96
1.6	15	5	98
0.8	8	6	99

ADJA-Paare (total 407 Paare)
(Mittelwert: 2.06 ±0.93)

in %	Anzahl	Distanz	kum.
70.8	288	2	71
17.2	70	1	88
6.9	28	3	95
2.2	9	5	97
1.7	7	4	99

ADJD-Paare (total 250 Paare)
(Mittelwert: 2.6 ±1.94)

in %	Anzahl	Distanz	kum.
66.0	165	2	66
10.8	27	3	77
10.0	25	1	87
4.4	11	4	91
2.4	6	5	94

TIGER

Alle Paare (total 2005 Paare)
(Mittelwert: 2.3 ±1.49)

in %	Anzahl	Distanz	kum.
61.4	1231	2	61
18.7	375	1	80
8.8	176	3	89
4.4	89	4	93
2.7	54	5	96
1.7	35	6	98
0.7	15	7	98
0.7	15	8	99

ADJA-Paare (total 1169 Paare)
(Mittelwert: 2.29 ±1.59)

in %	Anzahl	Distanz	kum.
53.0	619	2	53
24.7	289	1	78
9.4	110	3	87
5.1	60	4	92
3.7	43	5	96
1.9	22	6	98
0.9	10	7	99
0.8	9	8	100

ADJD-Paare (total 479 Paare)
(Mittelwert: 2.53 ±1.64)

in %	Anzahl	Distanz	kum.
61.8	296	2	62
12.5	60	3	74
12.1	58	1	86
4.8	23	4	91
2.7	13	6	94
2.3	11	5	96
1.3	6	8	98

Tabelle 4.5: Verteilung der paarweisen lexikalischen Token-Distanz zwischen den Köpfen von CAP in NEGRA und TIGER. Gezeigt werden die Fälle mit mindestens 5 Vorkommen.

NEGRA

Konj.	Anzahl	in %	Dist.	Anzahl	in %	kum.
1	926	96.8	2	792	82.8	83
			3	64	6.7	90
			5-9	36	3.8	93
			4	30	3.1	96
			10-	4	0.4	97
0	24	2.5	2	19	2.0	99
			1	3	0.3	99
			4	1	0.1	99
			3	1	0.1	99
3	4	0.4	5-9	2	0.2	100
			10-	2	0.2	100
2	3	0.3	5-9	2	0.2	100
			10-	1	0.1	100

TIGER

Konj.	Anzahl	in %	Dist.	Anzahl	in %	kum.
1	1949	97.2	2	1448	72.2	72
			3	207	10.3	82
			5-9	163	8.1	91
			4	123	6.1	97
			10-	8	0.4	97
0	30	1.5	2	19	0.9	98
			3	5	0.2	98
			5-9	3	0.1	98
			1	2	0.1	98
			4	1	0.0	98
2	22	1.1	5-9	17	0.8	99
			10-	3	0.1	99
			4	2	0.1	99
4	2	0.1	10-	2	0.1	100
3	2	0.1	5-9	1	0.0	100
			10-	1	0.0	100

Tabelle 4.6: Verhältnis der Konjunktoranzahl (Kommas und lexikalische Konjunktore) zur reinen Tokendistanz zwischen Köpfen von CAP in NEGRA und TIGER. Legende: Dist. = Tokendistanz(bereiche) inklusive Interpunktion; Konj. = Summe aller Kommas sowie der Token mit Wortart KON, welche zwischen den Köpfen erscheinen (nicht gezählt wurden Kommas, wenn sie direkt einem Konjunktor vorangehen). Lesebeispiel: Zwischen 36 Paaren von Köpfen aus benachbarten Konjunkttöchtern, welche 5-9 Token voneinander entfernt sind, ist nur ein Konjunktor vorhanden auf der Tokenebene.

NEGRA (total 550 Paare) (Mittelwert: 5.16 ±2.77)				TIGER (total 1141 Paare) (Mittelwert: 5.38 ±3.08)			
in %	Anzahl	Distanz	kum.	in %	Anzahl	Distanz	kum.
22.4	123	4	22	22.5	257	4	22
22.2	122	3	45	21.6	246	3	44
14.9	82	5	60	15.0	171	5	59
10.9	60	6	70	11.0	126	6	70
7.6	42	2	78	7.5	86	7	78
7.5	41	7	86	6.0	68	2	84
3.8	21	8	89	4.5	51	8	88
3.6	20	9	93	3.2	36	9	91
1.6	9	10	94	3.1	35	10	94
1.5	8	11	96	1.3	15	11	96
1.1	6	12	97	0.8	9	12	96
0.7	4	14	98	0.8	9	13	97
0.5	3	13	98	0.7	8	16	98
0.5	3	15	99	0.6	7	14	99
				0.4	4	15	99
				0.4	4	17	99
				0.3	3	20	100

Tabelle 4.7: Verteilung der paarweisen lexikalischen Token-Distanz der Köpfe von CPP in NEGRA und TIGER. Gezeigt werden die Fälle mit mindestens 3 Vorkommen.

CNP und CAP eher flach verteilt, gehen aber bis zu einem Maximum von 19 bei NEGRA und 26 bei TIGER.

4.1.3.1 Konjunktoranzahl zwischen Köpfen von CPP

Wie bei den CAP kann die Anzahl der Kommas und lexikalischen Konjunktoren (KON) ausgezählt werden, welche zwischen 2 Köpfen erscheinen. Da vor einigen Konjunktoren Komma stehen muss, wurden nur diejenigen Kommas gezählt, denen nicht unmittelbar ein lexikalischer Konjunktor nachfolgt. Die Tabellenzusammenstellung 4.8 auf der nächsten Seite zeigt die Verhältnisse für NEGRA und TIGER in Bezug auf die reine Tokendistanz. Nulldistanzen, d.h. weder Komma noch Konjunktor als Verknüpfungsmittel, sind bei den CPP insbesondere in NEGRA relativ verbreitet. Bis und mit Distanz 9 sind maximal 1 Koordinationsmittel zwischen den Konjunkten in der Mehrheit.

NEGRA

Konj.	Anzahl	in %	Dist.	Anzahl	in %	kum.
1	430	78.2	5-9	207	37.6	38
			4	110	20.0	58
			3	98	17.8	75
			10-	15	2.7	78
2	67	12.2	5-9	45	8.2	86
			10-	22	4.0	90
0	29	5.3	2	15	2.7	93
			3	8	1.5	94
			5-9	6	1.1	96
3	18	3.3	10-	12	2.2	98
			5-9	6	1.1	99
4	5	0.9	10-	5	0.9	100
13	1	0.2	10-	1	0.2	100

TIGER

Konj.	Anzahl	in %	Dist.	Anzahl	in %	kum.
1	908	79.6	5-9	413	36.2	36
			4	240	21.0	57
			3	204	17.9	75
			10-	49	4.3	79
			2	2	0.2	80
2	155	13.6	5-9	95	8.3	88
			10-	60	5.3	93
3	31	2.7	5-9	16	1.4	95
			10-	15	1.3	96
0	25	2.2	2	13	1.1	97
			5-9	5	0.4	97
			4	3	0.3	98
			3	3	0.3	98
			10-	1	0.1	98
4	15	1.3	10-	11	1.0	99
			5-9	4	0.4	100
5	4	0.4	10-	4	0.4	100
6	2	0.2	10-	2	0.2	100
8	1	0.1	10-	1	0.1	100

Tabelle 4.8: Verhältnis der Konjunktoranzahl (Kommas und lexikalische Konjunktore) zur reinen Tokendistanz zwischen Köpfen von CPP in NEGRA und TIGER. Legende: Dist. = Tokendistanz(bereiche) inklusive Interpunktion; Konj. = Summe aller Kommas sowie der Token mit Wortart KON, welche zwischen den Köpfen erscheinen (nicht gezählt wurden Kommas, wenn sie direkt einem Konjunktor vorangehen)

4.2 Morphologische Ähnlichkeit

Koordinierte Köpfe fungieren in syntaktisch gemeinsamen Umgebungen. Im Folgenden wird betrachtet, inwiefern sich dabei und daraus Ähnlichkeiten und Konsistenzen bezüglich morphologischer Kriterien auf der Ebene der Wortform und der Lemmata ergeben. Bei den flektierten Wortarten wäre zu erwarten, dass Flexionsendungen von Konjunkten eine stärkere Gemeinsamkeit aufweisen.

Weiter soll betrachtet werden, ob koordinierte Köpfe eine Tendenz zu gemeinsamen Derivationsendungen haben bzw. Endungsgleichklang (Homoeoteleuton) aufweisen. Im Folgenden wird dargestellt, inwieweit sich diese Vorstellung empirisch belegen lässt.

4.2.1 Übereinstimmung im Suffix bei CNP

Ein mögliches Kriterium für morphologische Ähnlichkeit auf der Ebene des Wortform ist die paarweise Übereinstimmung der Köpfe in einem Suffix der Länge 2. Wieviele der Köpfe haben überhaupt ein gemeinsames Suffix? In der Tabelle 4.9 auf der nächsten Seite sieht man, dass dieser Wert mit rund 20% ist nicht besonders hoch und erfolgversprechend ist.[5] Das häufigste gemeinsame Suffix ist die Flexionsendung „-en“, während „-ng“ und „-it“ typische Derivationssuffixe sind.

Um ein genaueres Bild über die Ähnlichkeit bezüglich gemeinsamer Derivationssuffixe zu bekommen, sind die Köpfe mit dem Morphologieanalyseprogramm GERTWOL (Haapalainen und Majorin 1994) lemmatisiert worden. Dieses Werkzeug berechnet ein flache morphotaktische Binnenstruktur, wo unter anderem starke Kompositionsgrenzen (#) mit allfällig vorangehendem Fugenelement (\), schwache Kompositionsgrenzen (|) und Grenzen für Derivationssuffixe (~) markiert sind. Für ein Wort wie „Derivationsgrenzen“ werden beispielsweise folgende Analysen berechnet:

```
"*deriv~ation\s#grenz~e"  S FEM PL NOM
"*deriv~ation\s#grenz~e"  S FEM PL AKK
"*deriv~ation\s#grenz~e"  S FEM PL DAT
"*deriv~ation\s#grenz~e"  S FEM PL GEN
"*deriv~ation\s#grenz~en"  S NEUTR SG NOM
"*deriv~ation\s#grenz~en"  S NEUTR SG AKK
"*deriv~ation\s#grenz~en"  S NEUTR SG DAT
```

Durch die hohe Abdeckung von GERTWOL entstehen oft Mehrfachanalysen auf der Ebene der Lemmata. Gemäss Volk (1999) entstehen über Zeitungstextkorpora etwa für 10% aller Substantive mehrdeutige Analysen. Wenn man starke Kompositagrenzen mit 4 Punkten, schwache mit 2 Punkten und Derivationsgrenzen mit 1 Punkt bestraft, ergibt sich durch die Addition aller Strafpunkte eines Lemmas

[5] Er liesse sich noch leicht erhöhen, wenn man nicht nur paarweise Köpfe von direkten Geschwisterkonjunkten vergleicht, sondern alle Köpfe innerhalb einer Koordination vergleicht.

NEGRA			
in %	Paare	Suffix	kum.
80.8	4894	—	81
8.7	525	en	90
2.4	147	er	92
1.8	107	ng	94
0.4	27	rn	94
0.4	26	ag	94
0.4	22	it	95
0.3	20	ße	95
0.2	15	es	95
0.2	15	in	96
0.2	14	st	96
0.2	13	ch	96
0.2	13	nd	96
0.2	12	im	96
0.2	11	te	97
0.1	9	ge	97
0.1	8	he	97
0.1	8	le	97
0.1	7	nt	97
0.1	6	de	97
0.1	6	el	97
0.1	6	ik	97
0.1	6	or	97
0.1	6	rk	98
0.1	6	se	98
0.1	6	us	98

TIGER			
in %	Paare	Suffix	kum.
79.9	8817	—	80
10.1	1116	en	90
2.6	290	ng	93
2.3	254	er	95
0.4	49	it	95
0.4	42	rn	96
0.3	38	te	96
0.2	26	ch	96
0.2	25	ns	96
0.2	21	nt	97
0.2	18	es	97
0.2	18	ie	97
0.2	17	in	97
0.1	16	on	97
0.1	13	ft	97
0.1	13	nd	98
0.1	13	us	98
0.1	12	ag	98
0.1	11	ik	98
0.1	11	rk	98
0.1	11	st	98
0.1	8	ät	98
0.1	8	ge	98
0.1	8	rs	98
0.1	7	s-	98
0.1	6	el	98
0.1	6	ic	99
0.1	6	n-	99

Tabelle 4.9: Verteilung der paarweise gemeinsamen Suffixe der Länge 2 in in Köpfen von CNP über NEGRA und TIGER. Gezeigt werden Mindestvorkommen ab 6.

NEGRA				TIGER			
in %	Paare	Suffix	kum.	in %	Paare	Suffix	kum.
92.4	5596	—	92	91.9	10143	—	92
2.1	126	ung	94	3.3	360	ung	95
1.2	75	e	96	1.3	140	er	96
1.2	74	er	97	0.8	92	e	97
0.3	21	en	97	0.3	32	en	98
0.2	10	straße	97	0.2	20	ie	98
0.1	6	Mark	98	0.1	16	Prozent	98
0.1	6	ist	98	0.1	16	keit	98
0.1	6	keit	98	0.1	10	Mark	98
0.1	5	Prozent	98	0.1	9	ik	98
0.1	5	chor	98	0.1	9	ität	98
0.1	5	in	98	0.1	7	ig	98
0.1	5	isch	98	0.1	7	in	98
0.1	5	ismus	98	0.1	7	ion	99
				0.1	6	isch	99
				0.1	6	s-	99
				0.0	5	ismus	99
				0.0	5	ist	99
				0.0	5	mann	99
				0.0	5	schlag	99

Tabelle 4.10: Verteilung der paarweise gemeinsamen lemmatisierten GERTWOL-Suffixe in CNP über NEGRA und TIGER. Das Suffix „—“ zeigt an, dass die beiden Köpfe unterschiedlich sind.

ein Präferenzmass, dass einfach strukturierte Lemmas bevorzugt. Wenn zusätzlich selten vorkommende Morpheme wie „stag“ oder „port“ bestraft werden, welche auffällig oft in mehrdeutigen Analysen erscheinen, lassen sich 90% der mehrdeutigen Lemmata korrekt auflösen (Volk 1999).

Für die Auswertung in der Tabelle 4.10 ist jeweils das letzte morphologische Segment eines Kopfes extrahiert worden. Es zeigt sich jedoch, dass der Anteil der paarweise gemeinsamen Suffixe durch die Lemmatisierung nicht zunimmt, sondern leider um mehr als die Hälfte abnimmt. Die Fälle der gemeinsamen Suffixe sind damit so selten, dass sich dieses Merkmal für praktische Zwecke kaum sinnvoll nutzen lässt.

4.2.2 Übereinstimmung im Suffix bei CAP

Attributive Adjektive Auf Grund der Kasusbedingungen wird für die ADJA-Kopfpaare eine starke Übereinstimmung bei den Köpfen erwartet. Um die Flexionsendungen der Adjektive zu erheben, wurden die Worformen mit einer einfachen

Regel reduziert: Nimm die letzten beiden Zeichen, falls das zweitletzte Zeichen ein „e“ ist, ansonsten nimm nur das letzte Zeichen. Damit werden auch Ordinalzahlen gleich gemacht, indem sie auf den Abkürzungspunkt reduziert werden.

Für NEGRA ergeben sich die folgenden Werte: „gleich“ (406, 99.8%), „unterschiedlich“ (1, 0.2%). Der Ausreisser hängt mit einem Rechtschreibefehler im Kontext von Strassennamen zusammen wie man im Beleg (198a) sieht. Ähnlich eindeutig sieht es in TIGER aus: „gleich“ (1164, 99.6%), „unterschiedlich“ (5, 0.4%). Neben Tokenisierungsfehlern tritt hier noch der Fall auf, wo eine Ordinalzahl in Ziffernschreibweise mit einem gewöhnlichen Adjektiv wie in (198b) auftritt.

(198) a. In unmittelbarer Nachbarschaft jenes umstrittenen Baus zwischen [$_{CAP}$ [$_{ADJA}$ Großem] und [$_{ADJA}$ Mittleren]] Hasenpfad, [...] [N_{13747}]

b. [...], wie sich die deutsche Seite die dauerhafte Einhaltung der Stabilitätskriterien in der [$_{CAP}$ 3. und letzten] Stufe der Europäischen Wirtschafts- und Währungsunion (EWWU) vorstellt [...] [T_{18831}]

Paarweises Vorkommen in Sätzen Wenn man im NEGRA-Korpus alle Paare von ADJA betrachtet, welche einander nachfolgen innerhalb eines Satzes, ergeben sich bezüglich Suffix-Gleichheit folgende Werte: „unterschiedlich“ (4584, 56.3%), „gleich“ (3551, 43.7%). Von den 3551 Paaren mit Suffix-Gleichheit sind allerdings nur knapp 336 koordiniert. Die Übereinstimmung für sich genommen ist somit kein diskriminierendes Merkmal bezüglich Koordinierheit. Nur wenn die Distanz zwischen den beiden Adjektiven 2 Token beträgt und gleichzeitig genau ein Komma oder ein Konjunktor dazwischen steht, ergibt sich ein eindeutiger CAP-Indikator, wie man der Tabellenzusammenstellung 4.11 auf der nächsten Seite insbesondere für das TIGER-Korpus entnehmen kann. Der etwas niedrigere Wert bei NEGRA hängt mit der nicht-konsequent durchgeführten CAP-Annotation bei asyndetischen Koordinationen zusammen. Die Tabelle zeigt, dass schon bei Distanzen ab 3 mehrheitlich keine CAP-Köpfe vorliegen.

Prädikative und adverbiale Adjektive Wie sieht die Übereinstimmung aus im Suffix der Länge 2, wenn jeweils zwischen zwei Köpfen mit dem STTS-Tag ADJD evaluiert wird? Wie die Auflistung (199) zeigt, haben im NEGRA-Korpus gut 21% dasselbe Suffix, beim TIGER-Korpus sogar knapp 27%.

(199) Verteilung der paarweisen Gleichheit im Suffix der Länge 2 bei ADJD in NEGRA:
„ungleich“ (196, 78.4%), „gleich“ (54, 21.6%)

(200) Verteilung der paarweisen Gleichheit im Suffix der Länge 2 bei ADJD in TIGER:
„ungleich“ (351, 73.1%), „gleich“ (129, 26.9%)

NEGRA				TIGER			
in %	Anzahl	Dist.	koord.	in %	Anzahl	Dist.	koord.
42.8	273	2	+	49.6	656	2	+
20.2	129	3	-	18.7	247	3	-
18.0	115	4	-	17.8	235	4	-
13.2	84	2	-	8.2	108	3	+
4.1	26	3	+	5.0	66	4	+
1.7	11	4	+	0.8	11	2	-

Tabelle 4.11: Verhältnis der Konjunktoranzahl (Kommas und lexikalische Konjunktore) zur reinen Tokendistanz zwischen Köpfen von CAP in NEGRA und TIGER.
Legende: Dist. = Tokendistanz(bereiche) (mit Interpunktion); Konj. = Summe aller Kommas sowie der Token mit Wortart KON, welche zwischen den Köpfen erscheinen (nicht gezählt wurden Kommas, wenn sie direkt einem Konjunktor vorangehen)

Die Auflistung (201) zeigt die mehr als einmal auftretenden Fälle der insgesamt 20 gemeinsamen Suffix-Types, wenn man das Material aus NEGRA und TIGER kombiniert.

(201) Liste der paarweise gleichen Suffixe der Länge 2 bei ADJD in TIGER und NEGRA mit mindestens 2 Vorkommen:
„ch“ (58, 31.7%), „er“ (49, 26.8%), „ig“ (17, 9.3%), „nd“ (17, 9.3%), „kt“ (6, 3.3%), „al“ (5, 2.7%), „ar“ (5, 2.7%), „it“ (4, 2.2%), „iv“ (4, 2.2%), „en“ (3, 1.6%), „ll“ (3, 1.6%), „am“ (2, 1.1%), „nt“ (2, 1.1%), „os“ (2, 1.1%)

Die Häufigkeiten der gleichen Endungspaare spiegeln dabei im Wesentlichen die Häufigkeiten der einzelnen Endungen in koordinierten CAP. Eine Ausnahme dazu ist die Endung „kt“, welche aus der Paarformel „direkt oder indirekt“ stammt.

(202) Liste der paarweise gleichen Suffixe der Länge 2 bei ADJD in TIGER und NEGRA mit mindestens 4 Vorkommen:
„ch“ (158, 21.6%), „ig“ (103, 14.1%), „er“ (66, 9.0%), „nd“ (59, 8.1%), „ar“ (29, 4.0%), „en“ (25, 3.4%), „rt“ (25, 3.4%), „ll“ (22, 3.0%), „al“ (20, 2.7%), „os“ (19, 2.6%), „iv“ (14, 1.9%), „et“ (13, 1.8%), „it“ (13, 1.8%), „nt“ (13, 1.8%), „lt“ (11, 1.5%), „kt“ (10, 1.4%), „ht“ (9, 1.2%), „am“ (8, 1.1%), „ft“ (8, 1.1%), „ng“ (7, 1.0%), „de“ (6, 0.8%), „el“ (6, 0.8%), „zt“ (6, 0.8%), „eu“ (4, 0.5%), „ös“ (4, 0.5%), „rn“ (4, 0.5%), „ut“ (4, 0.5%)

4.3 Semantische Ähnlichkeit

Die Köpfe von koordinierten Konjunkten stehen in einem gemeinsamen syntaktischen Verhältnis zu den Elementen, von denen sie abhängig sind. Im folgenden

Abschnitt soll untersucht werden, inwieweit damit auch eine semantische Nähe zwischen den Köpfen verbunden ist.

4.3.1 Hauptkategorien und Synonymmengen im GermaNet-Thesaurus

Eine einfache, aber relativ grobe Bestimmung von lexikalisch-semantischer Ähnlichkeit besteht darin, die Hauptkategorien von GermaNet 4.0[6] (Kunze 2005) zu verwenden. Jedes Synset, d.h. jede explizit[7] repräsentierte Bedeutung in GermaNet, ist genau einer Hauptkategorie[8] zugeordnet. Die folgenden Auflistungen zeigen die Besetzung dieser Hauptkategorien mit Synsets.

(203) Verteilung der total 27241 Nomen-Synsets auf die Hauptkategorien:
„Artefakt“ (3852, 14.1%), „Mensch“ (2702, 9.9%), „Geschehen“ (2698, 9.9%), „Kommunikation“ (2420, 8.9%), „Ort“ (2151, 7.9%), „Tier“ (2095, 7.7%), „Pflanze“ (1841, 6.8%), „Gruppe“ (1485, 5.5%), „Nahrung“ (1204, 4.4%), „Kognition“ (1001, 3.7%), „Koerper“ (940, 3.5%), „Substanz“ (854, 3.1%), „Menge“ (647, 2.4%), „Attribut“ (618, 2.3%), „Besitz“ (580, 2.1%), „Zeit“ (517, 1.9%), „natGegenstand“ (366, 1.3%), „natPhaenomen“ (364, 1.3%), „Gefuehl“ (352, 1.3%), „Relation“ (281, 1.0%), „Form“ (208, 0.8%), „Motiv“ (50, 0.2%), „Tops“ (15, 0.1%)

(204) Verteilung der total 1999 Adjektiv-Synsets auf die Hauptkategorien:
„Verhalten“ (284, 14.2%), „Pertonym“ (218, 10.9%), „Gefuehl“ (212, 10.6%), „Relation“ (205, 10.3%), „Koerper“ (157, 7.9%), „Zeit“ (147, 7.4%), „Perzeption“ (142, 7.1%), „Substanz“ (136, 6.8%), „Allgemein“ (124, 6.2%), „Gesellschaft“ (103, 5.2%), „Ort“ (94, 4.7%), „Geist“ (68, 3.4%), „natPhaenomen“ (48, 2.4%), „Menge“ (47, 2.4%), „Bewegung“ (14, 0.7%)

(205) Verteilung der total 8733 Verb-Synsets auf die Hauptkategorien:
„Veraenderung“ (1599, 18.3%), „Gesellschaft“ (1070, 12.3%), „Kognition“ (948, 10.9%), „Lokation“ (905, 10.4%), „Kommunikation“ (759, 8.7%), „Kontakt“ (482, 5.5%), „Besitz“ (466, 5.3%), „Koerperfunktion“ (463, 5.3%), „Schoepfung“ (434, 5.0%), „Allgemein“ (386, 4.4%), „Perzeption“ (354, 4.1%), „Gefuehl“ (329, 3.8%), „Konkurrenz“ (225, 2.6%), „Verbrauch“ (189, 2.2%), „natPhaenomen“ (124, 1.4%)

Das Einordnen der Wörter unter die Hauptkategorien ist teilweise überraschend. So sind etwa Eigennamen wie „Peugeot“ oder „Citroën“ in der Hauptkategorie

[6] Das grösste deutschsprachige Wortnetz ist unter `http://www.sfs.uni-tuebingen.de/GermaNet/` dokumentiert.

[7] Durch die Mehrfachvererbung in der Hyperonymie, welche in GermaNet häufiger benutzt wird als in WordNet, entstehen implizit bezüglich der Oberbegriffe weitere Bedeutungen.

[8] Die Schreibung der Hauptkategorien entspricht dem GermaNet-Original.

in %	Anzahl	Kategorie	+/-lemmatisierbar
65.2	33337	n	+
21.0	10723	v	+
7.2	3674	n	-
4.9	2487	a	+
1.6	813	v	-
0.2	89	a	-

Tabelle 4.12: Die Abdeckung der GermaNet-Lemmata durch das Morphologieanalysewerkzeug GERTWOL. Legende: n = Nomen, a = Adjektive, v = Verben

„Kognition“ zu finden, da sie unter der Hyperonymhierarchie „Automarke > Marke > Art > Kategorie > kognitives Objekt“ aufgeführt sind.

Die Anzahl Lemmata der drei Hauptwortarten Nomen, Adjektive, Verben sind in der Auflistung (206) gegeben:

(206) Verteilung der drei Hauptwortarten in GermaNet 4.0: Substantive (37576, 72.5%), Verben (11640, 22.5%), Adjektive (2589, 5.0%)

4.3.1.1 Lexikalische Abdeckung von GermaNet durch GERTWOL

Wieviele Wortformen aus NEGRA und TIGER sind in GermaNet repräsentiert? Um diese Frage beantworten zu können, müssen die flektierten Wortformen der Korpora mit den Lemmata von GermaNet vergleichbar gemacht werden. Zu diesem Zweck sind sowohl das GermaNet wie auch die Korpora mit Hilfe des Morphologieanalysewerkzeugs GERTWOL (Haapalainen und Majorin 1994) der Firma Lingsoft[9] lemmatisiert worden.

Über unrestringiertem Text, d.h. gemischtem Korpus mit literarischen Texten, Zeitungstexten und anwendungsspezifischen Texten wie Wetterberichten, erreicht GERTWOL gemäss Haapalainen und Majorin (1994) eine Abdeckung von 98%. Die Abdeckung von GERTWOL über den Lemmata von GermaNet ist in der Tabelle 4.12 zusammengestellt. Mit etwas über 90% Abdeckung ist dieser Wert deutlich tiefer als die von Lingsoft berechneten Zahlen. Da in einem Lexikon die flektierten Formen fehlen und GermaNet im Bestreben nach einer grossen semantischen Abdeckung innerhalb der bearbeiten Kategorien auch ungebräuchliche oder auch sehr fachsprachliche Lemmata aufführt, ist diese Abweichung nach unten zu erwarten. Der Einfluss von fachsprachlichen Ausdrücken[10] ist insbesondere bei den Substantiven stark spürbar. Daneben wirkt sich auch die Rechtschreibereform des Deutschen negativ aus, da GermaNet die neuen Konventionen verwendet, GERTWOL aber nicht.

[9]Eine Dokumentation der verwendeten Kategorien findet sich unter `http://www2.lingsoft.fi/doc/gertwol/intro/overview.html`.

[10]Prominent vertreten dabei sind Mineraliennamen wie Prasiolith, Achroit usw.

Obere Grenze der semantischen Abdeckung des NEGRA-Korpus durch GermaNet Wie sieht die Abdeckung von NEGRA und TIGER in den drei Haupt-Kategorien Nomen (Substantive und Eigennamen), Adjektiv und Verb, welche von GermaNet abgebildet werden, in GERTWOL aus? Bei den Nomen und Adjektiven können die Wortformen normalerweise direkt aus den Korpora an GERTWOL zur Analyse übergeben werden. Wenn GERTWOL eine Wortform nicht kennt, wird die Wortform selbst als Lemma genommen. Bei Bindestrichkomposita, welche bei Nomen oft Kombinationen aus Namen und normalen Substantiven darstellen, ist zusätzlich eine Reduktion auf das Letztglied hinter dem Bindestrich vorgenommen worden, da GermaNet selbst solche Kombination nur in geringem Mass enthält. Auf Grund der Rechtsköpfigkeit der deutschen Komposita stellt das Letztglied meist auch den semantischen Kern dar.

Im Fall der finiten Verben bilden die abtrennbaren Verbpräfixe als potentiell diskontinuierliche morphologische Elemente ein Problem. Im NEGRA-Korpus finden sich immerhin 2000 abgetrennte Verbpräfixe, was somit im Schnitt jeden zehnten Satz betrifft. Die Lemmata von finiten Verben, welche in der syntaktischen Struktur einen Knoten PTKVZ-SVP als Schwester besitzen, werden deshalb mit dem Lemma des Verpräfixes verkettet.

GermaNet enthält nicht nur Lemmata, welche durch eine einzelne Wortform realisiert werden, sondern weist innerhalb von Synsets auch mehrteilige Lemmata auf. Diese sind bei den normalen Substantiven typischerweise virtuelle, d.h. im Deutschen nicht lexikalisiert auftretende Einträge, welche der gleichmässigeren Oberbegriffsbildung dienen. In den folgenden Auswertungen sind jedoch nur Einwortlemmata berücksichtigt.

4.3.2 Semantische Ähnlichkeit bei CNP

Wieviele Substantive und Eigennamen (Vornamen und Beinamen sind in GermaNet nicht aufgeführt) aus untersuchten Korpora sind in GermaNet überhaupt erfasst?

4.3.2.1 Eigennamen

Bei dieser Kategorie gibt es oft Unsicherheit, was als echter Eigenname gelten soll und was nicht. Während für das Englische die Unterscheidung normalerweise am orthographischen Kriterium der Grossschreibung festgemacht wird, hilft dies im Deutschen nicht. Abgesehen von den eindeutigen Kategorien wie Personennamen und geographischen Bezeichnungen gibt es unterschiedliche Auffassungen, welche z.B. im STTS (Schiller u. a. 1999, 15) durch Angabe von Entitätsklassen (z.B. Planetennamen, Firmennamen) aufzählend spezifiziert sind. Es besteht zwar in GermaNet eine grosse Übereinstimmung mit den Konventionen von STTS, aber in einigen Punkten auch klare Abweichung. So sind Produktebezeichnungen in STTS als Gattungsnamen aufgefasst, in GermaNet erhalten sie jedoch Eigennamenmarkierung.

NEGRA (total 6166)				TIGER (total 9462)			
in %	Lemmata	Synsets	kum.	in %	Lemmata	Synsets	kum.
93.4	5757	0	93	93.5	8848	0	94
6.4	395	1	100	6.3	595	1	100
0.2	14	2	100	0.2	19	2	100

Tabelle 4.13: Verteilung der Type-Abdeckung von GermaNet bezüglich der Eigennamen in NEGRA und TIGER mit Reduktion der Bindestrichkomposita. In TIGER sind nur NE mitgezählt, die NNE nicht.

NEGRA (total 18958)				TIGER (total 41372)			
in %	Lemmata	Synsets	kum.	in %	Lemmata	Synsets	kum.
80.4	15245	0	80	73.8	30550	0	74
17.4	3306	1	98	24.5	10138	1	98
2.1	407	2	100	1.7	684	2	100

Tabelle 4.14: Verteilung der Token-Abdeckung von GermaNet bezüglich der Eigennamen in NEGRA und TIGER mit Reduktion der Bindestrichkomposita. In TIGER sind nur NE mitgezählt, die NNE nicht.

In GermaNet werden Organisationsbezeichnungen, insbesondere bei sogenannten „Unika“ (Bezeichnungen für kulturell als einmal vorkommend betrachtete Grössen) und wenn sie als Akronym realisiert sind, als Eigenname kodiert. Die nicht-abkürzende Form innerhalb desselben Synonymsets wird dann aber in einem gewissen Widerspruch dazu als Gattungsbezeichnung gehandhabt. So ist beispielsweise das Lemma „Bundesnachrichtendienst“ als normales Nomen kodiert, aber „BND“ als Eigenname. Darin kommt die Eigenschaft von Thesauri im Stile von WordNet zum Ausdruck, welche nicht explizit unterscheiden zwischen der Beziehung Gattung/Untergattung (*is-a)* sowie Gattung/Exemplar (*instance-of*) (Noy und McGuinness 2001).

Die Tabelle 4.13 zeigt die Abdeckung der verschiedenen Lemmata der Eigennamen (NE) in NEGRA und TIGER mit Reduktion der Bindestrichkomposita. Auf eine separate Auflistung der Resultate ohne Reduktion der Bindestrichkomposita wurde verzichtet, da die Unterschiede minim sind. Es sind nur etwas mehr als 6% der Lemma-Typen in GermaNet repräsentiert.

Mehrfache Sinne sind äusserst selten und umfassen Bezeichnungen wie „Wien“ oder „Washington“, welche in GermaNet die Hauptstadt und das Bundesland bezeichnen. Nicht damit kodiert sind metaphorische Verwendungen wie etwa bei Hauptstädten „Bonn“ oder „Berlin“, welche im politischen Tagesjournalismus gerne als Variante für die dort beheimatete Regierung oder Verwaltung verwendet werden.

Eine Auszählung über den einzelnen Vorkommen der Lemmata gibt die Ta-

belle 4.14 auf der vorherigen Seite. Da GermaNet insbesondere die häufigen und wichtigen Eigennamen enthalten sollte, ist eine höhere Abdeckung zu erwarten. In TIGER ist sie mit knapp 26% etwas höher als in NEGRA mit knapp 20%.

Paarweise Abdeckung von Kopfeigennamen in CNP In der Tabelle 4.15 auf der nächsten Seite ist dargestellt, wie oft bei zwei benachbarten NE-Köpfen eine gemeinsame Hauptkategorie gefunden wird in GermaNet. Die geographischen Bezeichnungen decken dabei die meisten Treffer ab, bei den Gruppen sind es vor allem Parteiabkürzungen wie „FDP“. Unter Kognition treten wie schon eingangs angesprochen vor allem Automarken auf.

Wenn man das kleinste gemeinsame Hyperonym (begrenzt auf maximal 3 Stufen der Hyperonymie) sucht zwischen den beiden Köpfen, ergeben sich abdeckungsmässig fast keine Unterschiede im Vergleich zu den Hauptkategorien. In der Tabelle 4.16 auf der nächsten Seite sieht man, dass damit jedoch semantisch besser fassbare Oberklassen entstehen. Die Idiosynkrasien von GermaNet sind allerdings beträchtlich. So sind einerseits detailliert Modellklassen von VW wie „Golf“ und „Polo“ aufgeführt (Hyperonym „VW $“), andrerseits ist „Hongkong“ nur als Stadtstaat und nicht als Stadt aufgeführt, sodass für die Koordination von „Bangkok“ und „Hongkong“ kein gemeinsames Hyperonym gefunden werden kann innerhalb der 3 Stufen.

4.3.2.2 Substantive

Die Tabellenaufstellung 4.17 auf Seite 226 zeigt für NEGRA die Abdeckung der insgesamt 21265 verschiedenen Lemmata von normalen Substantiven (NN), wenn Bindestrichkomposita nicht reduziert werden. Die Abdeckung von knapp 37% der Lemmata ist tief. Für TIGER, wo auch Eigennamenkomposita mit Substantiven als Letztglied (NNE) mit gezählt wurden, sind nur knapp 30% der total 31492 verschiedenen Lemmata abgedeckt.

Wenn man nicht die Types der Lemmata, sondern einzelnen Vorkommen zählt, wird die Abdeckung deutlich besser, da die häufig verwendeten Lemmata in GermaNet vollständiger repräsentiert sind. Die Tabellen in 4.18 auf Seite 226 geben die entsprechende Abdeckung an: Sowohl für TIGER wie NEGRA sind 75% der Vorkommen in GermaNet mit mindestens einem Synset abgedeckt.

Wenn die Reduktion der Bindestrichkomposita auf das Letztglied gemacht wird, zeigt sich eine Erhöhung der Abdeckung um rund 4%. Dies gilt sowohl bei der Auswertung auf der Token- wie auf der Type-Ebene, wie der Vergleich mit den Werten aus den Tabellen 4.19 und 4.20 auf Seite 227 ergibt.

Nur etwa 9% aller Lemmata haben mehr als eine Bedeutung innerhalb von GermaNet. Unter den Lemmata mit den meisten Synsets sind viele uninteressante länderspezifische Währungsvarianten aufgeführt, wie man der Auflistung (207) aus TIGER entnehmen kann. Unter den häufigen fehlenden Lemmata sind viele Abkürzungen und Akronyme zu finden.

NEGRA (total 737)				TIGER (total 1500)			
in %	Anzahl	Hauptkat.	kum.	in %	Anzahl	Hauptkat	kum.
61.2	451	—	61	54.7	821	—	55
33.1	244	Ort	94	41.3	620	Ort	96
5.2	38	Gruppe	100	3.5	52	Gruppe	100
0.5	4	Kognition	100	0.5	7	Kognition	100

Tabelle 4.15: Verteilung der paarweisen Übereinstimmung in der Hauptkategorie von GermaNet bezüglich der Eigennamen in NEGRA und TIGER. Um als Paar gezählt zu werden, müssen beide Köpfe das STTS-Tag NE aufweisen.

NEGRA (total 737)			TIGER (total 1500)		
in %	Anzahl	Hyperonym	in %	Anzahl	Hyperonym
61.3	452	—	55.9	838	—
20.2	149	Land	27.2	408	Land
5.4	40	Stadt	5.1	77	Stadt
3.7	27	Partei	2.9	44	Partei
2.2	16	Hauptstadt	2.6	39	Hauptstadt
1.5	11	Fernsehanstalt	1.8	27	Fläche
1.1	8	Fläche	0.9	14	Bundesland
0.9	7	Landeshauptstadt	0.5	8	Fluss
0.8	6	Bundesland	0.5	7	Fernsehanstalt
0.7	5	Kontinent	0.5	7	Landeshauptstadt
0.5	4	Automarke	0.4	6	Automarke
0.4	3	Fluss	0.3	4	Provinzhauptstadt
0.4	3	Verwaltungsgebiet	0.2	3	Insel
0.4	3	geographisches Gebiet	0.2	3	Kontinent
0.1	1	Kanton	0.2	3	geographisches Gebiet
0.1	1	Provinzhauptstadt	0.1	2	Inselstaat
0.1	1	Region	0.1	2	Region
			0.1	2	Regionshauptstadt
			0.1	2	Verwaltungsgebiet
			0.1	1	Bundesstaat
			0.1	1	Inselgruppe
			0.1	1	Union
			0.1	1	VW $

Tabelle 4.16: Verteilung der paarweisen Übereinstimmung des niedrigsten Hyperonyms bezüglich der Eigennamen in NEGRA und TIGER. Es wurden nur Hyperonyme gezählt, welche maximal über 3 direkte Hyperonym-Relationen vermittelt sind. Lesebeispiel: In NEGRA haben in 8 Fällen die beiden NE-Köpfe von 2 benachbarten Konjunkten einer CNP als niedrigstes gemeinsames Hyperonym das Synset „Fläche“.

NEGRA (total 21265)				TIGER (total 31492)			
in %	Lemmata	Synsets	kum.	in %	Lemmata	Synsets	kum.
62.7	13331	0	63	70.4	22158	0	70
29.5	6278	1	92	24.0	7559	1	94
5.5	1179	2	98	4.1	1284	2	98
1.5	324	3	99	1.1	336	3	100
0.5	102	4	100	0.3	103	4	100
0.1	28	5	100	0.1	27	5	100
0.1	14	6	100	0.0	14	6	100
0.0	5	7	100	0.0	6	7	100
0.0	1	8	100	0.0	2	8	100
0.0	1	9	100	0.0	1	9	100
0.0	1	10	100	0.0	1	10	100
0.0	1	15	100	0.0	1	15	100

Tabelle 4.17: Verteilung der Type-Abdeckung von GermaNet bezüglich der Substantive in NEGRA und TIGER ohne Reduktion von Bindestrichkomposita. In TIGER sind auch die NNE mitgezählt.

NEGRA (total 73730)				TIGER (total 147359)			
in %	Lemmata	Synsets	kum.	in %	Lemmata	Synsets	kum.
43.3	31926	1	43	44.1	64952	1	44
25.1	18480	0	68	24.0	35419	0	68
16.4	12073	2	85	16.6	24457	2	85
7.7	5709	3	92	8.4	12405	3	93
4.8	3537	4	97	4.4	6537	4	98
1.3	922	5	99	1.2	1706	6	99
1.2	852	6	100	0.9	1350	5	100
0.2	151	7	100	0.2	236	7	100
0.0	36	15	100	0.1	192	15	100
0.0	29	8	100	0.0	59	8	100
0.0	10	9	100	0.0	24	10	100
0.0	5	10	100	0.0	22	9	100

Tabelle 4.18: Verteilung der Token-Abdeckung von GermaNet bezüglich der Substantive in NEGRA und TIGER ohne Reduktion von Bindestrichkomposita. In TIGER sind auch die NNE mitgezählt.

NEGRA (total 19209)				TIGER (total 27994)			
in %	Lemmata	Synsets	kum.	in %	Lemmata	Synsets	kum.
58.2	11172	0	58	66.3	18562	0	66
33.1	6363	1	91	27.3	7648	1	94
6.2	1196	2	98	4.6	1290	2	98
1.7	324	3	99	1.2	338	3	99
0.5	103	4	100	0.4	104	4	100
0.1	28	5	100	0.1	27	5	100
0.1	14	6	100	0.1	14	6	100
0.0	5	7	100	0.0	6	7	100
0.0	1	8	100	0.0	2	8	100
0.0	1	9	100	0.0	1	9	100
0.0	1	10	100	0.0	1	10	100
0.0	1	15	100	0.0	1	15	100

Tabelle 4.19: Verteilung der Type-Abdeckung von GermaNet bezüglich der Substantive in NEGRA und TIGER mit Reduktion von Bindestrichkomposita. In TIGER sind auch die NNE mitgezählt.

NEGRA (total 73730)				TIGER (total 147359)			
in %	Lemmata	Synsets	kum.	in %	Lemmata	Synsets	kum.
45.2	33299	1	45	45.9	67660	1	46
21.7	16001	0	67	21.0	30885	0	67
17.2	12684	2	84	17.3	25477	2	84
8.1	6000	3	92	8.8	12904	3	93
5.0	3651	4	97	4.6	6729	4	98
1.3	963	5	98	1.2	1764	6	99
1.2	896	6	100	0.9	1390	5	100
0.2	156	7	100	0.2	240	7	100
0.0	36	15	100	0.1	203	15	100
0.0	29	8	100	0.0	61	8	100
0.0	10	9	100	0.0	24	10	100
0.0	5	10	100	0.0	22	9	100

Tabelle 4.20: Verteilung der Token-Abdeckung von GermaNet bezüglich der Substantive in NEGRA und TIGER mit Reduktion von Bindestrichkomposita. In TIGER sind auch die NNE mitgezählt.

(207) Lemmata aus TIGER mit mindestens 6 verschiedenen Synsets in GermaNet:
15 (Dollar), 10 (Franc), 9 (Krone), 8 (Dinar), 8 (Verbindung), 7 (Abschluß), 7 (Band), 7 (Peso), 7 (Pfund), 7 (Punkt), 7 (Ring), 6 (Abgang), 6 (Absatz), 6 (Berufung), 6 (Bild), 6 (Geschichte), 6 (Klasse), 6 (Land), 6 (Programm), 6 (Raum), 6 (Rupie), 6 (Spur), 6 (Teil), 6 (Übertragung), 6 (Zug)

Durch die Reduktion der Bindestrichkomposita gibt es mehr identische Köpfe innerhalb eines Paares; für NEGRA total 58 und für TIGER total 103 . In den nachfolgenden semantischen Auswertungen sind die Fälle mit den identischen Köpfen jeweils ausgeblendet.

(208) Die häufigsten identischen Kopf-Substantive aus NEGRA und TIGER: „Prozent“ (21, 13.0%), „Mark“ (16, 9.9%), „Dr.“ (6, 3.7%), „S.“ (4, 2.5%), „Gramm“ (3, 1.9%), „Jahr“ (3, 1.9%), „Jahren“ (3, 1.9%), „MGC“ (3, 1.9%), „SV“ (3, 1.9%), „Seite“ (3, 1.9%)

Paarweise Abdeckung von Kopf-Substantiven in CNP In der Tabelle 4.21 auf der nächsten Seite ist dargestellt, wie oft bei zwei benachbarten NN-Köpfen eine gemeinsame Hauptkategorie gefunden wird in GermaNet. Die Abdeckung liegt bei 64%, was etwa 2% über der erwarteten kombinierten Wahrscheinlichkeit von $0.79\% \times 0.79\% \approx 0.62\%$ liegt. Die Auftretenshäufigkeit der gefundenen Hauptkategorien entspricht sich in beiden Korpora bis auf wenige Ausnahmen.

Wenn man analog zu den Eigennamen die kleinsten gemeinsamen Hyperonyme innerhalb von maximal 3 Stufen berechnet, zeigt sich bei den Substantiven ein anderes Bild. Die feinere Strukturierung mit den konzeptuellen Zwischenklassen (erkennbar am „?“) ergibt eine viel höhere Zahl unterschiedlicher Hyperonyme (TIGER: 370 , NEGRA: 305). Die Auflistung in (209) zeigt, dass zudem die Abdeckung für TIGER um knapp 12% sinkt.

(209) Kleinste gemeinsame Hyperonyme (maximal 3 Stufen) von NN-Köpfen in TIGER (mindestens 10 Vorkommen):
„—“ (5871, 76.9%), „Mensch“ (127, 1.7%), „Handlung“ (109, 1.4%), „Berufstätiger“ (72, 0.9%), „Geschehen“ (51, 0.7%), „Gefühl“ (42, 0.6%), „Agens ?“ (33, 0.4%), „Artefakt“ (32, 0.4%), „Organisation“ (30, 0.4%), „staatliche Institution“ (29, 0.4%), „Veränderung“ (27, 0.4%), „Gruppe“ (25, 0.3%), „Qualität“ (25, 0.3%), „Himmelsrichtung“ (22, 0.3%), „Zeiteinheit“ (21, 0.3%), „Kontinentalbürger ?“ (20, 0.3%), „Mitmensch“ (19, 0.2%), „Zusammenschluss“ (19, 0.2%), „Handlungsweise“ (17, 0.2%), „Monat“ (17, 0.2%), „Objekt“ (17, 0.2%), „Stelle“ (17, 0.2%), „Europäer“ (16, 0.2%), „Verwaltungsgebiet“ (16, 0.2%), „Verwandter“ (16, 0.2%), „gleichrangiger Mensch ?“ (15, 0.2%), „wirtschaftliche Institution“ (15, 0.2%), „Arbeit“ (14, 0.2%), „Maßeinheit“ (14, 0.2%), „Asiat“ (13, 0.2%), „Attribut“ (13, 0.2%), „Beziehung“ (13, 0.2%), „Mitteilung“

NEGRA (total 4375)			TIGER (total 7632)		
in %	Anzahl	Hauptkategorie	in %	Anzahl	Hauptkategorie
64.1	2805	—	65.1	4965	—
10.6	465	Mensch	10.0	763	Mensch
4.8	210	Geschehen	5.8	442	Geschehen
4.6	200	Artefakt	3.7	281	Gruppe
3.6	156	Gruppe	3.3	250	Artefakt
1.9	81	Kognition	1.9	147	Kognition
1.8	77	Ort	1.7	133	Ort
1.6	69	Nahrung	1.7	127	Kommunikation
1.3	58	Kommunikation	1.1	83	Zeit
1.2	51	Zeit	0.9	71	Besitz
0.8	36	Gefuehl	0.9	65	Attribut
0.6	28	Koerper	0.7	55	Gefuehl
0.6	27	Besitz	0.6	45	Koerper
0.5	23	Attribut	0.5	39	Nahrung
0.5	21	Tier	0.5	36	Menge
0.4	18	Pflanze	0.3	25	natPhaenomen
0.4	18	Substanz	0.3	24	Pflanze
0.4	17	Menge	0.3	24	Substanz
0.2	8	Relation	0.3	23	Relation
0.1	5	natPhaenomen	0.3	22	Tier
0.0	2	natGegenstand	0.1	5	Form
			0.1	4	Motiv
			0.0	3	natGegenstand

Tabelle 4.21: Verteilung der paarweisen Übereinstimmung in der Hauptkategorie von GermaNet bezüglich der Substantive in NEGRA und TIGER. Um als Paar gezählt zu werden, müssen beide Köpfe das STTS-Tag NN aufweisen.

(13, 0.2%), „Regierungsbeamter“ (13, 0.2%), „öffentliche Institution“ (13, 0.2%), „Wissenschaftler“ (11, 0.1%), „Produkt“ (10, 0.1%), „Vorgesetzter“ (10, 0.1%), „soziale Gruppe“ (10, 0.1%)

(210) Kleinste gemeinsame Hyperonyme (maximal 3 Stufen) von NN-Köpfen in NEGRA (mindestens 10 Vorkommen):
„—“ (3292, 75.2%), „Mensch“ (86, 2.0%), „Handlung“ (34, 0.8%), „Gefühl“ (28, 0.6%), „Berufstätiger“ (26, 0.6%), „Artefakt“ (25, 0.6%), „Mitmensch“ (25, 0.6%), „Geschehen“ (21, 0.5%), „Verwandter“ (21, 0.5%), „Stelle“ (18, 0.4%), „Lerner“ (17, 0.4%), „Verwaltungsgebiet“ (16, 0.4%), „Gebäude“ (15, 0.3%), „Agens ?“ (14, 0.3%), „Kind“ (14, 0.3%), „Gruppe“ (13, 0.3%), „staatliche Institution“ (12, 0.3%), „Himmelsrichtung“ (11, 0.3%), „Künstler“ (11, 0.3%), „Organisation“ (11, 0.3%), „Qualität“ (11, 0.3%), „Zusammenschluss“ (11, 0.3%), „örtlich bestimmter Mensch ?“ (10, 0.2%)

Für beide Korpora ergeben sich insgesamt folgende Abdeckungswerte: „-“ (9163, 76.3%), „+“ (2844, 23.7%)

Evaluation der Hyperonym-Bestimmung Um die Korrektheit der Hyperonym-Bestimmung von GermaNet zu bestimmen, wurde intellektuell evaluiert. Anhand der Cochran-Formel Bartlett u. a. (2001) wurde für die knapp 2850 Fälle aus TIGER und NEGRA eine Stichprobengrösse von 250 bestimmt, sodass mit 90% Wahrscheinlichkeit die Gesamtkorrektheit innerhalb $\pm 5\%$ des Evaluationsresultates liegt. Bei den abstrakteren Hauptkategorien ist eine Beurteilung manchmal unscharf. Es wurde anhand der Frage „Kann man x in diesem Kontext als eine Art z auffassen?“ eher grosszügig entschieden. Systematische Mehrdeutigkeit wie Gebäude und Institution bei „Kindergarten“ oder „Schule“ wurden in beiden Lesarten als korrekt beurteilt.

Die Resultate in (211) ergeben eine hohe Korrektheit der zugeordneten Hyperonyme.

(211) Evaluation von 250 zufällig ausgewählten Stichproben aus TIGER und NEGRA: „Korrekt“ (241, 96.4%), „Falsch“ (9, 3.6%)

Synonymie-Relation Echte standardsprachliche Synonyme wie „Beginn“ und „Anfang“ sind wenig häufig in GermaNet. Regionale Varianten wie „Rahm“ und „Sahne“ oder „Marillenschnaps“ und „Aprikosenschnaps“ machen nebst reinen orthographischen Varianten wie „Brokkoli“ und „Broccoli“ einen grossen Teil der in GermaNet ausgewiesenen Synonymiebeziehungen aus.

Daneben gibt es bei den Nomen über 2250 morphologisch derivierte Geschlechtervarianten vom Typ „Erzieher“ vs. „Erzieherin“[11], welche im Gegensatz zu mor-

[11] Es gibt auch noch 5 Geschlechtervarianten vom Typ „Westdeutsche“ vs. „Westdeutscher“. Diese Bildungen kommen nur im Zusammenhang mit herkunftsbezogenen Personenbezeichnungen mit

phologisch unabhängigen Bezeichnungen wie „Nichte“ vs. „Neffe“, „Bruder“ vs. „Schwester“ jeweils zusammen in einem Synset vereinigt sind.

GermaNet-Synonymie zwischen Köpfen von CNP tritt in NEGRA nur in 86 Fällen (2%) und in TIGER sogar nur in 99 Fällen (1.3%) auf. Die Auflistung (212) dokumentiert die typischen Vertreter, die sich vor allem aus Kombinationen von weiblichen und männlichen Anredevarianten und Paarformeln wie „Art und Weise“ nähren. Die Koordination von echten Synonymen wäre auch redundant und ein Verstoss gegen Konversationsmaximen (Grice (1975)).

(212) Auflistung der 97 verschiedenen synonymen NN-Köpfe in CNP aus TIGER und NEGRA (mindestens 3 Vorkommen):
„Bürgerinnen Bürger“ (11, 5.9%), „Schülerinnen Schüler“ (11, 5.9%), „Art Weise“ (10, 5.4%), „Kolleginnen Kollegen“ (10, 5.4%), „Wählerinnen Wähler“ (8, 4.3%), „Forschung Entwicklung“ (6, 3.2%), „Ort Stelle“ (6, 3.2%), „Transport Verkehr“ (6, 3.2%), „Bürger Bürgerinnen“ (4, 2.2%), „Gewerkschafterinnen Gewerkschafter“ (4, 2.2%), „Seniorinnen Senioren“ (4, 2.2%), „Genossinnen Genossen“ (3, 1.6%), „Grund Boden“ (3, 1.6%), „Hohn Spott“ (3, 1.6%), „Spielerinnen Spieler“ (3, 1.6%), „Teilnehmerinnen Teilnehmer“ (3, 1.6%)

Es gibt zwar eine starke Tendenz zur festen Reihenfolge in Paarformeln[12] – z.B. im Fall von männlich-weiblich Alternation gemäss Höflichkeitskonvention die Voranstellung der weiblichen Form – Abweichungen sind aber trotzdem anzutreffen. Beim häufigsten Synonym-Paar im NEGRA-Korpus, „Schüler/Schülerin“, ist in 3 von 14 Fällen die männliche Form an erster Stelle.

Die stehende Wendung „Art und Weise“ ist das häufigste koordinierte Synonympaar, das keine Geschlechteralternativen ausdrückt. Die starke Lexikalisierung dieser Koordination zeigt sich auch in den Suchresultaten grosser Suchmaschinen. Während „Art und Weise“ in deutschsprachigen Texten bei Google im September 2005 gut 9 Millionen Mal gefunden wird, sind es bei „Weise und Art“ nur knapp 400 Treffer. Die Kombination „Grund Boden“ wurde in Form von „Boden und Grund“ nur etwa 500 Mal verzeichnet, in der andern Reihenfolge gab es knapp 900000 Treffer.

Antonymie-Relation Noch seltener mit insgesamt 53 Vorkommen in NEGRA und TIGER ist die Antonymie, welche in GermaNet als lexikalische semantische Relation auf der Ebene der Lemmata kodiert ist.

(213) Auflistung der 34 verschiedenen antonymen NN-Köpfe in CNP aus TIGER und NEGRA (mindestens 2 Vorkommen):

dem Suffix „deutsche/r“ vor und werden auch nur bei diesen für beide Geschlechter aufgeführt. Andere deadjektivische Nominalisierungen wie „Angestellter“ vs „Angestellte“ werden nur in der männlichen Form aufgelistet.

[12] Vgl. Müller (1997), welcher verschiedene Präferenzen optimalitätstheoretisch aufarbeitet.

„Ende Anfang“ (5), „Eltern Kinder“ (4), „Schüler Lehrer“ (4), „Arbeitgebern Arbeitnehmern“ (3), „Arbeitgeber Arbeitnehmer“ (3), „Anfang Ende“ (2), „Arbeitnehmern Arbeitgebern“ (2), „Erfolg Mißerfolg“ (2), „Kinder Erwachsene“ (2), „Verkäufer Käufer“ (2)

4.3.3 Semantische Ähnlichkeit bei CAP

Die Tabellenzusammenstellung 4.23 auf der nächsten Seite zeigt die Abdeckung der verschiedenen Adjektiv-Lemmata (ADJA sowie ADJD) ohne Berücksichtigung von Ordinalzahladjektiven in Ziffernschreibweise. Die Abdeckung ist mit etwa 20% bei NEGRA und etwa 16% bei TIGER markant schlechter als bei den Substantiven. Die Reduktion von Bindestrichkomposita bei Adjektiven bringt nur eine leichte Verbesserung um 0.6%, wie in der Tabelle 4.25 auf Seite 234 ersichtlich.

Die Abdeckung auf der Ebene der Token-Vorkommen ist mit gut 55% bei beiden Korpora etwa gleich. Die Tabelle 4.25 auf Seite 234 zeigt, dass auch hier die Lemmata mit mehr als einem Synset öfter vorkommen als Adjektive mit einem Synset.

Die häufigsten fehlenden Lemmata sind in der Auflistung (214) auf dieser Seite zu sehen.

(214) Häufigste Adjektiv-Lemmata aus TIGER, welche nicht in GermaNet repräsentiert sind (ab Mindestvorkommen 50)[13]:
„ander“ (652), „erst“ (610), „eigen“ (434), „weiter“ (391), „international“ (341), „letzt“ (271), „geplant“ (180), „zweit“ (169), „kommend“ (167), „israelisch“ (158), „französisch“ (144), „offenbar“ (133), „zusätzlich“ (128), „sogenannt“ (105), „dritt“ (102), „kroatisch“ (101), „laufend“ (99), „zunehmend“ (92), „gesellschaftlich“ (91), „umstritten“ (88), „stellvertretend“ (83), „nigerianisch“ (82), „führend“ (78), „bosnisch“ (75), „besonder“ (71), „türkisch“ (69), „vereint“ (67), „bayerisch“ (66), „weitgehend“ (65), „offensichtlich“ (62), „ostdeutsch“ (60), „umgerechnet“ (59), „japanisch“ (59), „recht“ (56), „wachsend“ (55), „palästinensisch“ (55), „serbisch“ (52), „dänisch“ (52), „niederländisch“ (51), „Münchner“ (51)

Dafür gibt es verschiedene Gründe:

- Generell sind in GermaNet relativ wenige Adjektive aufgeführt und ordinale Adjektive fehlen fast ganz.
- GermaNet verzeichnet departizipiale Adjektive nicht. Attributiv verwendete Wortformen wie „überstandene“ oder „schreitende“ sind deshalb nicht abgedeckt.
- Zusammengesetzte Adjektive wie „staatskritisch“ fehlen meist, ausser wenn sie eine starke Lexikalisierung aufweisen wie „althergebracht“.

[13] Im Fall von „französisch“ liegt ein Tippfehler in GermaNet vor. Es steht dort „französich“.

NEGRA (total 5427)				TIGER (total 8152)			
in %	Lemmata	Synsets	kum.	in %	Lemmata	Synsets	kum.
79.7	4323	0	80	84.3	6870	0	84
16.9	919	1	97	13.3	1088	1	98
2.7	147	2	99	1.9	154	2	100
0.5	26	3	100	0.3	27	3	100
0.1	7	4	100	0.1	7	4	100
0.1	3	5	100	0.0	4	5	100
0.0	1	8	100	0.0	1	8	100
0.0	1	10	100	0.0	1	10	100

Tabelle 4.22: Verteilung der Type-Abdeckung von GermaNet bezüglich der Adjektive in NEGRA und TIGER ohne Reduktion von Bindestrichkomposita. Es wurden alle Lemmata mit dem STTS-Tag ADJA und ADJD ausgewertet ausser den Ordinalzahlen in Ziffernschreibweise.

NEGRA (total 26548)				TIGER (total 58045)			
in %	Lemmata	Synsets	kum.	in %	Lemmata	Synsets	kum.
44.4	11778	0	44	44.7	25959	0	45
37.2	9878	1	82	36.9	21445	1	82
12.2	3227	2	94	12.4	7175	2	94
4.3	1145	3	98	4.5	2624	3	98
1.4	383	5	100	1.0	584	5	100
0.2	65	4	100	0.2	108	4	100
0.1	38	8	100	0.2	96	10	100
0.1	34	10	100	0.1	54	8	100

Tabelle 4.23: Verteilung der Token-Abdeckung von GermaNet bezüglich der Adjektive in NEGRA und TIGER ohne Reduktion von Bindestrichkomposita. Es wurden alle Lemmata mit dem STTS-Tag ADJA und ADJD ausgewertet ausser den Ordinalzahlen in Ziffernschreibweise.

NEGRA (total 5331)				TIGER (total 7941)			
in %	Lemmata	Synsets	kum.	in %	Lemmata	Synsets	kum.
79.3	4225	0	79	83.9	6659	0	84
17.3	921	1	97	13.7	1088	1	98
2.8	147	2	99	1.9	154	2	100
0.5	26	3	100	0.3	27	3	100
0.1	7	4	100	0.1	7	4	100
0.1	3	5	100	0.1	4	5	100
0.0	1	8	100	0.0	1	8	100
0.0	1	10	100	0.0	1	10	100

Tabelle 4.24: Verteilung der Type-Abdeckung von GermaNet bezüglich der Adjektive in NEGRA und TIGER mit Reduktion von Bindestrichkomposita. Es wurden alle Lemmata mit dem STTS-Tag ADJA und ADJD ausgewertet ausser den Ordinalzahlen in Ziffernschreibweise.

NEGRA (total 26548)				TIGER (total 58045)			
in %	Lemmata	Synsets	kum.	in %	Lemmata	Synsets	kum.
44.0	11675	0	44	44.3	25732	0	44
37.4	9941	1	81	37.3	21622	1	82
12.3	3265	2	94	12.4	7225	2	94
4.3	1145	3	98	4.5	2624	3	98
1.4	383	5	99	1.0	584	5	100
0.3	67	4	100	0.2	108	4	100
0.1	38	8	100	0.2	96	10	100
0.1	34	10	100	0.1	54	8	100

Tabelle 4.25: Verteilung der Token-Abdeckung von GermaNet bezüglich der Adjektive in NEGRA und TIGER mit Reduktion von Bindestrichkomposita. Es wurden alle Lemmata mit dem STTS-Tag ADJA und ADJD ausgewertet ausser den Ordinalzahlen in Ziffernschreibweise.

NEGRA (total 3585)				TIGER (total 4600)			
in %	Lemmata	Synsets	kum.	in %	Lemmata	Synsets	kum.
38.7	1387	1	39	39.7	1827	1	40
22.6	812	0	61	27.2	1250	0	67
19.8	709	2	81	17.8	818	2	85
8.1	292	3	89	6.7	308	3	91
4.8	171	4	94	3.9	180	4	95
2.1	76	5	96	1.6	75	5	97
1.4	51	6	98	1.1	52	6	98
0.8	28	7	98	0.6	29	7	99
0.5	17	8	99	0.4	18	8	99
0.4	14	9	99	0.3	14	9	99
0.2	7	11	99	0.2	7	11	100
0.1	5	12	100	0.1	5	12	100
0.1	4	15	100	0.1	4	15	100
0.1	4	10	100	0.1	4	10	100
0.1	3	13	100	0.1	3	13	100

Tabelle 4.26: Verteilung der Type-Abdeckung von GermaNet bezüglich der Verben in NEGRA und TIGER.

- Adjektive mit Ziffernbestandteilen wie „27jährige“ sind nicht in GermaNet. Es gibt auch keine Metanotation für solche Wortbildungen.

4.3.4 Semantische Ähnlichkeit bei CVP

Die Tabelle 4.26 zeigt die Abdeckung der verschiedenen Lemmata von Verben mit den STTS-Tags VVFIN, VVINF, VVIMP und VVPP. Die Abdeckung ist bei dieser Kategorie mit über 77% bei NEGRA und knapp 73% bei TIGER am höchsten. Auch der Anteil von gegen 20% der Lemmata mit 2 Synsets ist deutlich höher als bei den bisher betrachteten Wortarten. Andererseits sind die Verblemmata von den drei Hauptkategorien die kleinste Gruppe, d.h. die lexikalische Varianz ist hier viel kleiner.

Auf der Ebene der Token betrachtet ergeben sich Abdeckungsgrade von knapp 96% bei TIGER und NEGRA, womit sich der auf der Ebene der Lemmatypen sichtbare Unterschied von 4% einebnet. Die exakten Verteilungen für die beiden Kopora enthält die Tabelle 4.27 auf der nächsten Seite.

Die Auflistung (215) der häufigsten fehlenden Abdeckungen in NEGRA zeigt, dass die Fehler teilweise im Korpus selbst (ein falsch tokenisiertes HTML-Aufzählungszeichen „bullet“ als VVPP) liegen, teilweise auch mit GERTWOL zusammenhängen („gegenüberstehen“, „handele“) oder tatsächlich nicht in GermaNet vorhanden sind („vorschweben“). Verben mit abtrennbaren Präfixen sind besonders betroffen.

NEGRA (total 26116)				TIGER (total 53495)			
in %	Lemmata	Synsets	kum.	in %	Lemmata	Synsets	kum.
19.0	4959	1	19	19.8	10570	1	20
18.4	4817	3	37	17.9	9585	3	38
17.0	4438	2	54	17.0	9072	2	55
12.4	3234	4	67	12.8	6837	4	68
8.5	2228	5	75	8.4	4517	5	76
5.6	1474	6	81	5.6	2980	6	82
4.2	1107	0	85	4.1	2207	0	86
3.2	838	7	88	3.4	1804	7	89
2.2	585	15	90	2.1	1126	8	91
2.0	511	8	92	1.8	978	11	93
1.8	461	11	94	1.8	971	15	95
1.4	361	18	96	1.1	596	18	96
1.0	259	9	97	1.0	509	9	97
0.8	206	12	98	0.8	415	12	98
0.7	183	14	98	0.6	344	26	98
0.5	141	26	99	0.6	332	14	99
0.5	124	10	99	0.5	278	10	99
0.4	104	13	100	0.4	208	13	100
0.3	86	17	100	0.3	166	17	100

Tabelle 4.27: Verteilung der Token-Abdeckung von GermaNet bezüglich der Verben in NEGRA und TIGER.

(215) „zustandekommen“ (10), „vorschweben“ (8), „gegenüberstehen“ (8), „bullet“ (8), „hinzukommen“ (7), „hervorgehen“ (7), „widerspiegeln“ (6), „reisten“ (6), „abwandern“ (6), „zugutekommen“ (5), „nachrücken“ (5), „konzertieren“ (5), „brachliegen“ (5), „absichern“ (5), „vorrechnen“ (4), „vgl.“ (4), „umwandeln“ (4), „schwerfallen“ (4), „hindeuten“ (4), „hervorheben“ (4), „handele“ (4), „einspeisen“ (4), „einhergehen“ (4), „bereithalten“ (4), „bedingt“ (4), „zurückstufen“ (3), „zurufen“ (3), „zuneigen“ (3), „verdichten“ (3), „unterkriegen“ (3), „reinkommen“ (3), „ordern“ (3), „mißfallen“ (3), „hochrechnen“ (3), „hinauskommen“ (3), „hereinholen“ (3), „großschreiben“ (3), „gegenüberstellen“ (3), „freikommen“ (3), „entgegenbringen“ (3), „einläuten“ (3), „ausschildern“ (3), „ausliefern“ (3), „abgewinnen“ (3)

Fazit Ingesamt lässt sich festhalten, dass die obere Grenze der semantischen Abdeckung durch GermaNet über den 3 Hauptkategorien unterschiedlich ausfällt. Insbesondere die Adjektive, aber auch die normalen Substantive weisen grosse Lücken bezüglich lexikalischem Material von NEGRA auf.

4.4 Koordinierendes Komma

In der Orthographie des Standarddeutschen wird das Komma als satzinternes Anknüpfungszeichen verwendet, das unterschiedliche Funktionen wahrnimmt. Es trennt Nebensätze von Hauptsätzen ab, markiert grössere Infinitiv- und Partizipialgruppen, herausgehobene Satzteile wie Anreden und Interjektionen sowie Zusätze und Einschübe. Daneben spielt es bei den reihenden und koordinativen Verknüpfungen eine wichtige Rolle.

4.4.1 Experimente zum Erkennen von koordinierendem Komma

Für das Kategorisieren von Kommata in koordinierende bzw. nicht-koordinierende Exemplare wurden verschiedene Experimente mit Hilfe von supervisierten maschinellen Lernverfahren durchgeführt.

4.4.1.1 Aufbereitung der Lerndaten

Da die Kommata in den verwendeten Korpora nicht explizit in ihrer syntaktischen Funktion spezifiziert und nicht in die syntaktische Struktur eingebettet sind, muss diese Information aus dem strukturellen Kontext hochgerechnet werden. Als satz- und phraseninternes Interpunktionsmittel kann dies in kontinuierlichen Phrasen zuverlässig gemacht werden nach folgendem Algorithmus:

1. Berechne für jede Phrase die unmittelbaren Konstituenten in der Struktur.

2. Suche für jedes Komma die am tiefsten liegende Konstituente, welche sowohl das am nächsten links und rechts vom Komma stehende Terminal dominiert, das in die Struktur integriert ist.

3. Falls diese Konstituente eine koordinierte Kategorie vom Typ *K* ist, dann kategorisiere das Komma als \$,*K*.

Probleme Bei diskontinuierlichen Konstituenten kann dieses Verfahren dazu führen, dass die Kommata etwas zu hoch eingehängt werden: Der Syntaxgraph in Abbildung 4.2 auf der nächsten Seite illustriert dies – sowie die Tatsache, dass neben dem Komma noch weitere nicht in die Struktur integrierte Elemente stehen können.

Dieser Satz zeigt zudem, dass nicht bloss Satzendpunkte in der Rechtschreibung zusammengezogen werden (nämlich bei Abkürzungen am Satzende), sondern dass dies auch bei Kommata der Fall sein kann. So vereinigt das Komma nach der eingebetteten Redeeinleitung einerseits die Abgrenzung der Redeeinleitung zur direkten Rede, andrerseits trennt es die beiden koordinierten Teilsätze ab.

Bei Fällen von koordinierten Sätzen wie in Beispiel (216a) auf dieser Seite, wo ein lexikalischer Konjunktor mit vorangehendem Komma auftritt, stellt sich die Frage, ob man das Komma als koordinierend betrachten möchte oder nicht. Man mag auf den ersten Blick nur die Funktion der Teilsatzabgrenzung darin sehen. Wenn man das analoge Beispiel (216b) dazu betrachtet, ist die Setzung des Kommas doch wieder stark an den Konjunktor geknüpft, bzw. an die mit dem adversativen „sondern“ spezifisch verknüpfte Sprechbetonung und -pause. Die Frage der Kommasetzung in teilsatzverknüpfenden Kontexten war in der amtlichen Regelung von 1901/1902 nicht reglementiert und das durch den Rechtschreibe-Duden vorgegebene komplexe Regelsystem dazu wurde mit der Rechtschreibereform (Zwischenstaatliche Kommission für Deutsche Rechtschreibung 2005) etwas relaxiert.[14]

(216) a. Die folkloristischen Elemente beschränken sich nicht allein auf Brasilien **, sondern** lassen auch asiatische und afrikanische Akzente einfließen. [N_{20}]

b. Die folkloristischen Elemente beschränken sich auf Brasilien **und** lassen keine asiatische und afrikanische Akzente einfließen.

Für die nachfolgenden Experimente wurden alle Kommata, welche vor einem Konjunktor stehen, als koordinierend markiert.

[14]Grundsätzlich wird in §71 festgehalten, dass koordinierte Strukturen durch Komma getrennt sind: „Gleichrangige (nebengeordnete) Teilsätze, Wortgruppen oder Wörter grenzt man mit Komma voneinander ab.“ Die konjunktorspezifischen Regelungen zur Kommasetzung sind in §72 wie folgt aufgeführt: „Sind die gleichrangigen Teilsätze, Wortgruppen oder Wörter durch *und, oder, beziehungsweise/bzw., sowie* (= und), *wie* (= und), *entweder . . . oder*, *nicht . . . noch, sowohl. . . als (auch)*, *sowohl . . . wie (auch)* oder durch *weder . . . noch* verbunden, so setzt man kein Komma.“

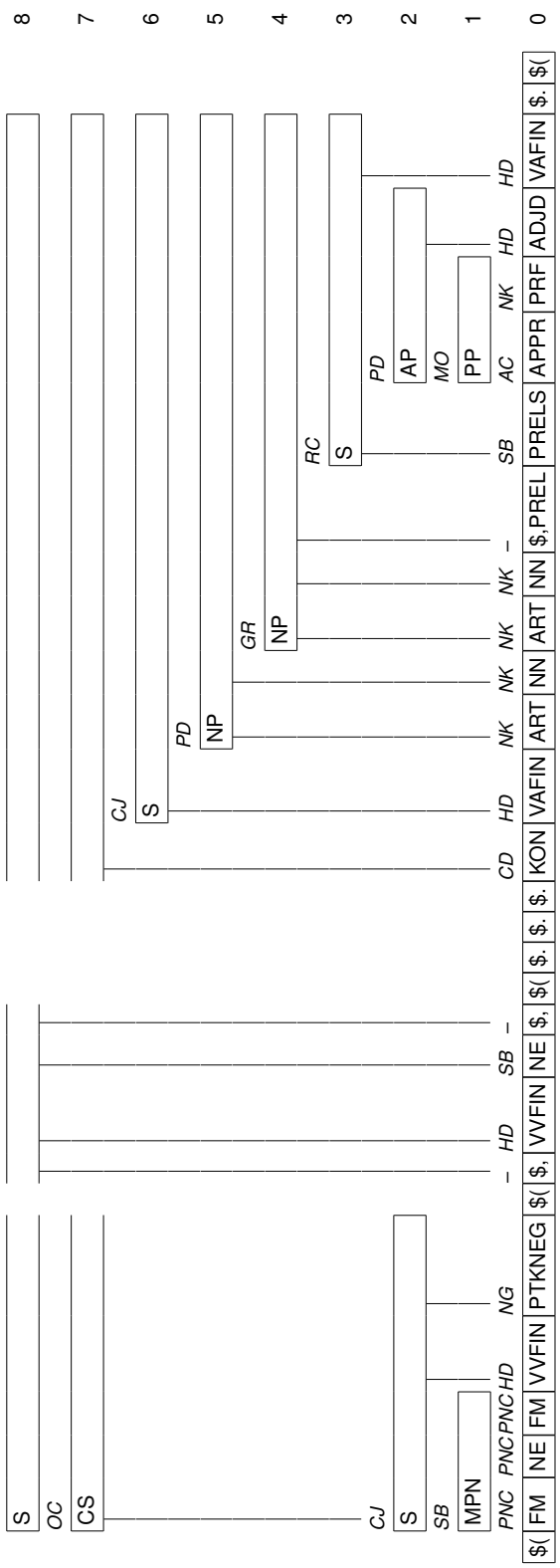


Abbildung 4.2: Satz 42 aus NEGRA mit einem Komma, das gleichzeitig koordinierte Teilsätze trennt und der Abgrenzung einer eingebetteten Redeeinleitung dient

4.4.2 Automatische Klassifikation von koordinierenden Kommata

Mit Hilfe des frei verfügbaren Programmes `megam`[15], das ein schnelles Lernverfahren für Modelle mit maximaler Entropie realisiert, wurde mit der automatischen Klassifikation der Funktion von Kommata experimentiert. Eine Einführung in die Sprachmodellierung mit Hilfe dieses Ansatzes gibt Berger u. a. (1996), eine Sammlung von praktischen Anwendungen für Desambiguierungsprobleme auf verschiedenen Stufen der Sprachverarbeitung enthält Ratnaparkhi (1998).

4.4.2.1 Multiklassen-Klassifikation über Wortarten-Tags

Zum Lernen wurden die aufbereiteten Daten aus dem NEGRA-, TIGER- und CZ-Korpus verwendet. Damit die Resultate einem realistischen Szenario entsprechen, sind beim Testen nicht die perfekten, d.h. manuell korrigierten Wortarten, verwendet worden. Die Testdaten sind automatisch vom entscheidungsbaumbasierten Wortarten-Tagger `TreeTagger 3.1` (Schmid 1995) mit der darin enthaltenen grossen Standard-Parameterdatei ohne zusätzliche lexikalische Ressourcen getaggt worden.

Die STTS-basierten Wortartentags dieser Parameterdatei sind fast vollständig kompatibel mit den Wortartentags in den verwendeten Korpora. Einige der bestehenden Unterschiede können durch einfaches Abbilden eingeebnet werden: Für Pronominaladverbien (im TIGER-Korpus immerhin mit knapp 4000 Vorkommen vertreten) wird in der Parameterdatei das „veraltete“ STTS-Tag PAV verwendet, in den Korpora jedoch das neuere Kürzel PROAV. Bei den attribuierenden Indefinitpronomen wird beim TIGER-Korpus nicht zwischen PIDAT und PIAT unterschieden – es wird nur noch die Kategorie PIAT ausgezeichnet. Daneben gibt es noch Unterschiede bei der Behandlung von APPR vs. KOKOM, welche sich schlecht automatisch auflösen lassen ausser durch die Vereinigung der beiden Kategorien[16]. Verfeinerte Tags innerhalb der Korpora wie die im CZ-Korpus mit INTADR annotierten Internet-Adressen oder die im TIGER-Korpus mit NNE annotierten Komposita, welche aus Eigennamen und normalen Nomen zusammengesetzt sind, sind vom Tagger als NN klassifiziert.

Die globale Tagging-Genauigkeit auf der Ebene der Wortarten ergibt dabei mit den Standard-Einstellungen für die 3 Korpora Korrektheitswerte um 95% wie in der Tabelle 4.28 auf der nächsten Seite ersichtlich.

Koordinierende Kommata Die Tabelle 4.29 zeigt die Verteilung aller koordinierten Kommata in den verwendeten Korpora. Knapp ein Viertel aller Kommata ist demzufolge koordinierend. Vorkommen von weniger als 500 sind in den folgen-

[15] Das Programm ist verfügbar unter `http://hal3.name/megam`. Details zur Implementation des Lernverfahrens, das schneller als das traditionell verwendete GIS-Verfahren (*generalized iterative scaling*) ist, finden sich im unpublizierten Artikel „Notes on CG and LM-BFGS Optimization of Logistic Regression“ unter `http://pub.hal3.name#daume04cg-bfgs`.

[16] Vgl. Teufel (1995) zum Problem des Abgleichs von Tagsets.

Korpus	Token	Fehler	Genauigkeit in %
NEGRA	355028	17329	95.12
TIGER	712332	34689	95.13
CZ	62381	3725	94.03

Tabelle 4.28: Übersicht zu Grösse und Fehlerquote der vom `TreeTagger` getaggten Korpora

in %	Anzahl	Tag	kumulativ
77.8	43716	$,	77.8
9.9	5588	$,CNP	87.7
8.2	4592	$,CS	95.9
1.5	868	$,CAP	97.4
1.2	647	$,CPP	98.6
1.0	548	$,CVP	99.6
0.4	219	$,CO	100.0
0.0	21	$,CAVP	100.0
0.0	3	$,CAC	100.0
0.0	1	$,CVZ	100.0

Tabelle 4.29: Verteilung der Funktionen der total 56203 Kommata in allen verwendeten Korpora

den Resultaten nicht dargestellt, weil das Lernverfahren bei so wenig Trainingsmaterial keine vernünftigen Ergebnisse ergibt.

4.4.2.2 Evaluationsresultate

Die Abbildung 4.3 auf Seite 243 zeigt die Lernkurven bezüglich F-Mass für das Optimieren über einem Kontext der Wortarten von 5 Token links und rechts vom Komma über den automatisch getaggten Texten. Es wurde in Lernschritten von 5000 Token trainiert und 10-fach kreuzvalidiert. Die Standardabweichungen sind teilweise beträchtlich, wie man der Darstellung 4.4 auf Seite 243 sehen kann.

Die häufigsten Kategorien werden tendenziell am zuverlässigsten erkannt, nur die zweithäufigste Koordinations-Kategorie $,CS fällt heraus. Eine Kategorisierung mit lokalen und zudem strukturell kaum abstrahierenden Merkmalen stösst an ihre Grenzen, denn zur korrekten Klassifikation müssten grössere Einheiten betrachtet werden und strukturelle Information zur Verfügung stehen. Die Standardabweichung bei $,CS, wie in Abbildung 4.4 auf Seite 243 ersichtlich, ist trotz der bescheidenen Erkennungsrate (gemessen an den zur Verfügung stehenden Trainingsexemplaren) recht klein; die Klassifikation ist mit etwas über 45% F-Mass zwar eher schlecht, dies aber stabil.

Auffällig ist die Kurve für CAP, welche zunächst einen steilen Lernzuwachs hat, um dann ab 35000 Komma-Token plötzlich wieder stark abzusinken. In der

Abbildung 4.4 auf der nächsten Seite, welche die Standardabweichung mitverzeichnet, zeigt sich, dass die Verschlechterung der Resultate mit einer Erhöhung der Standardabweichung einhergeht. D.h. die Trainingsdaten konvergieren nicht gut. Ruhiger verhält sich die CNP-Kurve, welche durch einen konstanten, aber langsamen Anstieg bis auf knapp 80% kommt.

Wenn man den Ausschnitt aus der Konfusionsmatrix in Abbildung 4.5 auf Seite 244 betrachtet, sieht man, dass $,CAP in erster Linie mit dem nicht-koordinierenden $, verwechselt wird. Eine qualitative Analyse der Falsch-Kategorisierungen zeigt, dass diese oft problematisch sind, zumindest in dem betrachteten Ausschnitt, der aus dem TIGER-Korpus stammt. Im Beispiel (217a) wurde maschinell beim fett ausgezeichneten Komma ein $,CAP kategorisiert. Die TIGER-Annotation mit 3 NP macht für mich wenig Sinn, tatsächlich bildet „repräsentativ“ zusammen mit „staatstragend“ eine CAP. In Beispiel (217b) sieht das maschinelle Lernverfahren ebenfalls eine CAP mit koordinierendem Komma, die Annotation ist aber entgegen den eigenen Annotationskonventionen ohne CAP gemacht. Im Beispiel (217c) liegt die Ursache bei einem Tagging-Fehler, welcher „gelähmt“ als VVPP kategorisiert.

(217) a. Doch nie ging es Yun um [$_{CNP}$ [$_{NP}$ die pompöse] , [$_{NP}$ die repräsentative] **,** [$_{NP}$ [$_{AP}$ gar „ staatstragende “] Wirkung]] seiner Musik. [T_{6051}]

b. [$_{PP}$ Nach einer [$_{ADJA-NK}$ stundenlangen] **,** [$_{AP-NK}$ von großem Ernst geprägten] Diskussion] [. . .] [T_{5904}]

c. Deshalb sind sie [$_{CAP}$ so erschüttert , entsetzt **,** wie gelähmt]. [T_{5734}]

Training des Komma-Klassifikators über automatisch getaggtem Korpus Um den Einfluss der Tagging-Fehler zu verkleinern, wurde in einem zusätzlichen Experiment nicht mit den perfekten N-Grammen von Wortarten aus dem Goldstandard trainiert, sondern mit den fehlerbehafteten automatisch getaggten N-Grammen. Die Kommas darin wurden dann mit der korrekten Klassifikation aus dem Goldstandard für das Training versehen.

Die Hypothese, die in diesem Experiment zu prüfen war, lautete: Da das automatische Tagging Fehler produziert, kann es sinnvoller sein, das optimale Modell mit diesen Fehlern zu berechnen, um dem Modell die Möglichkeit zu geben, die Tagging-Fehler über die Goldstandard-Klassifikation mitzukorrigieren.

Wie die Abbildung 4.6 auf Seite 244 bzw. die Tabelle 4.30 deutlich zeigt, gelingt es `megam` aber nicht, die fehlerbehafteten Tags zur besseren Klassifikation über fehlerbehaftetem Testmaterial zu benutzen. Die Leistung des Multiklassen-Klassifikators geht insgesamt zurück. Besonders zerstörerisch ist der Effekt bei Kommas in CVP und CPP, den beiden seltensten Komma-Klassen in dieser Auswertung.

Multiklassen-Klassifikation vs. multiple Binomialklassifikation In einem weiteren Experiment wurde getestet, ob die Leistung besser wird, wenn anstelle eines einzigen Klassifikators, der die verschiedenen Klassen zuweisen kann, für jede

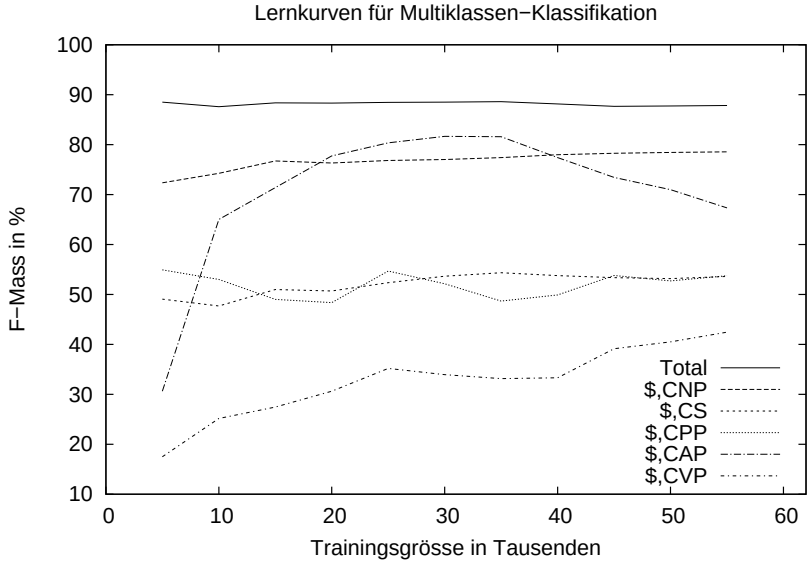


Abbildung 4.3: Lernkurve der Multiklassen-Klassifikation der Kommas 10-fach kreuzvalidiert für supervisiertes Lernen mit `megam`. Die Trainingsmenge wurde jeweils um 5000 Exemplare erhöht. Die mit „Total“ beschriftete Kurve umfasst auch alle nicht-koordinierenden Kommas, welche 75% der Fälle stellen und dementsprechend gut erkannt werden.

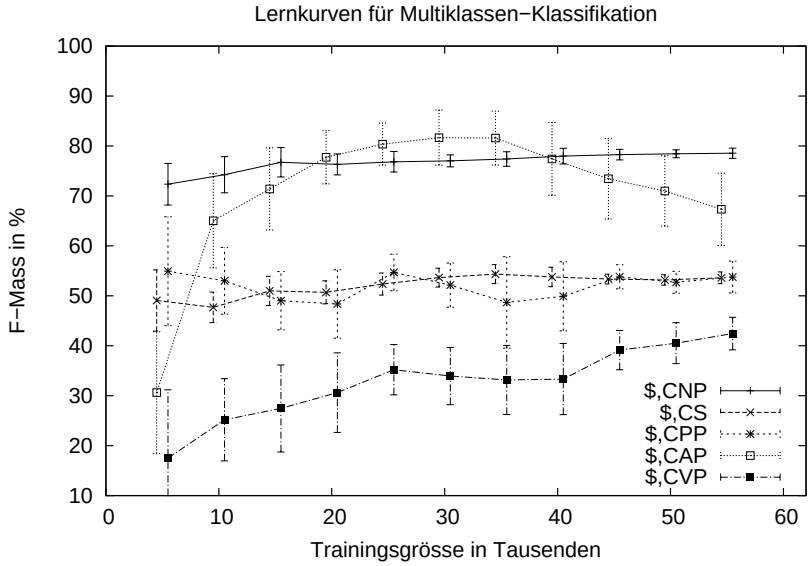


Abbildung 4.4: Lernkurve der Multiklassen-Klassifikation 10-fach kreuzvalidiert für supervisiertes Lernen mit `megam` für ausgewählte Kommatypen mit Standardabweichungen. Um die Lesbarkeit zu erhöhen, wurden die Datenpunkte leicht verschoben auf der x-Achse.

confusion matrix			error ratio	
wrong tag	correct tag	frequency	rel ct	total
$,CAP	$,	17	0.58	0.30
$,	$,CAP	11	23.91	0.20
$,CS	$,CAP	3	6.52	0.05
$,CAP	$,CPP	2	2.50	0.04
$,CNP	$,CAP	2	4.35	0.04
$,CAP	$,CNP	2	0.49	0.04
$,CAP	$,CS	2	0.48	0.04
...				
all	all	578		10.28

Abbildung 4.5: Ausschnitt aus der Konfusionsmatrix des Testsets eines Lerndurchgangs mit den Fehlern zu $,CAP. Legende: rel ct = Relativer Fehler bezüglich korrektem Tag

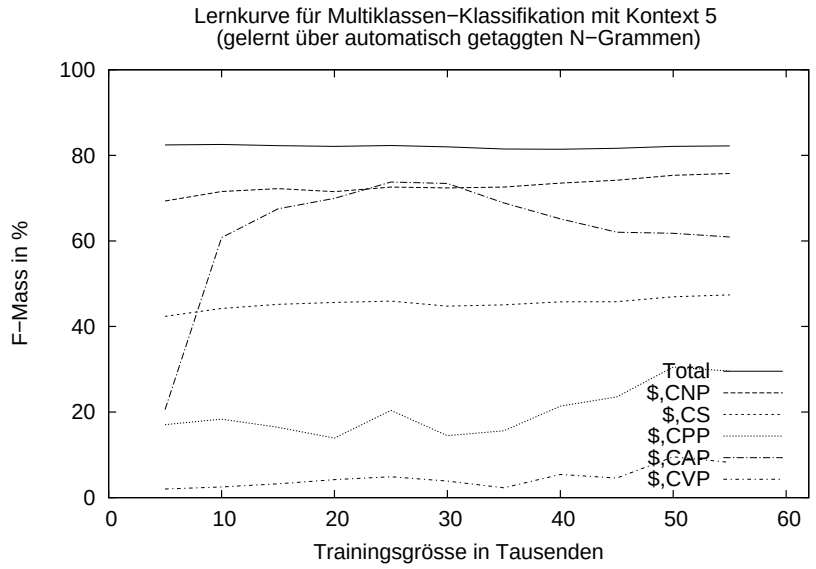


Abbildung 4.6: Lernkurve 10-fach kreuzvalidiert für supervisiertes Lernen mit `megam` mit automatisch getaggten N-Grammen gelernt

POS-Tags	$,CNP	$,CS	$,CPP	$,CAP	$,CVP	Total
Perfekt	78.5	53.6	53.8	67.3	42.4	87.8
Automatisch	75.8	47.4	29.6	60.9	8.2	82.2

Tabelle 4.30: Klassifikationsresultate (F-Mass) für das Training über perfekten und automatisch berechneten POS-Tags über der grössten Trainingsmenge.

Klasse separat ein binärer Klassifikator trainiert wird. Die Resultate liessen sich damit jedoch nicht verbessern.

4.4.2.3 Ausblick

Für eine Verbesserung der Kategorisierung von Kommas können verschiedene Ansätze weiter verfolgt werden. Anstelle der in den Experimenten verwendeten vektorbasierten Darstellung des Kontexts lassen sich im *Maximum-Entropy*-Ansatz flexibel klassenspezifisch Merkmale erheben und verrechnen. Alternativ dazu ist die Verwendung von anderen in der Computerlinguistik erfolgreich eingesetzten maschinellen Lernverfahren wie dem Ansatz der *Conditional Random Fields* (Lafferty u. a. 2001) oder dem *Memory-Based Learning* (Daelemans und van den Bosch 2005) denkbar.

Eine weitere Idee besteht darin, für die nicht koordinierenden Verwendungen der Kommas feinere Unterscheidungen zu treffen (Kommas bei Relativsätzen, bei Nebensätzen, welche mit KOUS eingeleitet werden, usw.). Brants (1997) hat gezeigt, dass die Verwendung von feineren internen Tagsets einen positiven Effekt auf die Gesamtleistung haben kann.

Kapitel 5

Koordonymie

Die automatische Gewinnung von lexikalisch-semantischen Taxonomierelationen aus Textkorpora hat in der Sprachverabeitung eine längere Tradition. Seit den Arbeiten von Hearst (1992) und Iwanska u. a. (2000) gehört die Verwendung von syntaktischen Mustern („Hearst-Patterns“), welche insbesondere auch Koordinationsphänomene zu diesem Zweck ausnützen, ins Repertoire.

5.1 Koordonymie als korpusbasierte Kopfrelation

Zwei Lemmata x und y sind genau dann **koordonym** (notiert als $coord(x,y)$) in einem Korpus C, wenn gilt: Es gibt in C mindestens eine koordinierte Struktur mit mindestens zwei Konjunkten, deren Köpfe aus x bzw. y bestehen. Diese Relation ist symmetrisch, d.h. wenn $coord(x,y)$ gilt, dann gilt auch $coord(y,x)$.

Zwei Lemmata x und y sind genau dann **paarweise koordonym** in einem Korpus C, wenn gilt: Es gibt in C mindestens eine koordinierte Struktur mit mindestens zwei adjazenten Konjunkten, deren Köpfe aus y bzw. x bestehen.

Ein **Koordonymset** eines Lemmas x ist die Menge aller Lemmata y, welche mit x koordonym sind.

$$coordset(x) = \{y \mid coord(x,y)\}$$

5.2 Koordonymie bei Nomen

5.2.1 Syntaktische Arten von Koordonymie

Die semantische Nähe von koordinierten Nomen hängt unter anderem davon ab, wie nah und unmodifiziert syntaktisch Konjunkte stehen, ob attributive Attribute oder postnominale Modifikatoren die einzelnen Konjunkte bedeutungsmässig spezifizieren oder abwandeln. Mit der strukturellen Information lassen sich syntaktische Unterklassen von Koordonymie zu betrachten.

5.2.2 Syntaktische Konjunkttypen

Die Köpfe kooordinierter Phrasen werden unterschiedlich modifiziert. Im Folgenden werden drei Typen von Konjunkten anhand ihrer Modifikation unterschieden:

- Ein Konjunkt vom **Typ A** besteht aus den Kernen von Wortkoordination und Wortgruppenkoordinationen, welche höchstens durch Artikel, attribuierende Pronomen oder Zahlen, aber nicht durch attributive Adjektive erweitert sind.
- Ein Konjunkt vom **Typ B** besteht aus den Kernen von Wortgruppen- und Phrasenkoordinationen, welche höchstens um attributive Adjektive oder Adjektivphrasen erweitert sind.
- Ein Konjunkt vom **Typ C** besteht aus Kernen von Wortgruppen und Phrasenkoordinationen, welche postnominal erweitert sind durch Genitiv-Attribute, Relativsätze oder postnominale Präpositionalphrasen.

Die Appositionen und Parenthesen werden bewusst ignoriert, da ich davon ausgehe, dass sie im Gegensatz zu den andern Arten von Modifikation eher ein kleines semantisches Veränderungspotiential haben. Ausser Betracht gelassen werden auch alle Modifikationen, welche die Koordination als Ganzes betreffen.

Typ A ist mit Beispiel (218a) illustriert, ein Konjunkt vom Typ B erscheint in (218b) und ein Konjunkt vom Typ C in (218c).

(218) a. Auch Philip Glass wurde auf seinen weltweiten Tourneen mit [$_{CNP}$ **Kassetten** und **Tonbändern**] überschüttet. [N_9]

b. In den Fragen etwa [$_{CNP}$ der **inneren Sicherheit** oder der Ausländerpolitik] sind die Positionen der Republikaner, der NPD und der CDU nicht sehr weit auseinander. [N_{146}]

c. Nachdem sich sein übereifriger Sportkumpel beharrlich weigerte, dem Gerempelten [$_{CNP}$ Schmerzensgeld und einen **Ausgleich für den erlittenen Verdienstausfall**] zu zahlen, zog der Kicker vor Gericht - und verlor. [T_{260}]

Wenn man die syntaktische Typbestimmung über den drei Korpora für CNP[1] macht, ergeben sich die Zahlen in Tabelle 5.1 auf der nächsten Seite. Wie aus den Resultaten der strukturellen Untersuchungen im Kapitel 2.2 zu erwarten, sind in allen drei Korpora Konjunkte vom Typ A mit rund 80% dominant.

Die Klassifikation der Konjunkte kann nun benutzt werden, um die paarweisen Koordonymiebeziehungen in den 9 möglichen Kombinationstypen aufzuschlüsseln. Die Tabelle 5.3 auf Seite 251 zeigt die Verteilung der Varianten über alle Kombinationen. Erwartungsgemäss dominieren die AA-Paare. C-Konjunkte treten deutlich häufiger an 2. Stelle auf, es sei denn, dass die beiden Konjunkte parallel gebaut erscheinen. Beispiel (219a) illustriert den Kombinationstyp CC, (219b) den Typ AB und (219c) den Typ BB.

[1] Mehrteilige Eigennamen werden nicht berücksichtigt.

in %	Anzahl	Konjunkttyp
NEGRA		
79.4	9136	A
10.6	1225	B
10.0	1152	C
TIGER		
75.3	15440	A
13.0	2660	C
11.7	2396	B
CZ		
80.8	2116	A
11.6	304	C
7.6	200	B

Tabelle 5.1: Verteilung der Konjunkttypen A, B und C in NEGRA, TIGER und CZ

(219) a. Die EU bietet den überwiegend ehemaligen britischen und französischen Kolonien [$_{CNP}$ freien Marktzugang für Industriegüter und Präferenzen für Agrarprodukte] an. [T_{4534}]

b. [. . .] , wenn während des Urlaubs mit der Vermittlung [$_{CNP}$ eines Arbeitsplatzes oder einer beruflichen Qualifizierung] zu rechnen ist . [T_{2772}]

c. Die übrigen 750 000 Landwirte fürchten neben der EG-Agrarpolitik , die ihnen [$_{CNP}$ niedrigere Subventionen und damit geringere garantierte Abnahmepreise] verordnen will, [. . .] [T_{605}]

5.2.3 GermaNet-Hauptkategorien in Kombinationstypen

In diesem Abschnitt wird an Hand der Abdeckung von GermaNet-Hauptkategorien untersucht, ob sich die Köpfe der unterschiedlichen Kombinationstypen bezüglich ihrer semantischen Nähe unterscheiden. Die Tabelle 5.3 auf Seite 251 gibt für die 9 Kombinationstypen die entsprechende Aufschlüsselung. Die relativen Anteile sind in der Auflistung (220) gegeben und zeigen, dass die symmetrischen Kombinationstypen am die höheren Übereinstimmunganteile aufweisen, was als Indikator für eine stärkere semantische Nähe gewertet werden kann.

(220) Relative Anteile der von GermaNet abgedeckten Hauptkategorien pro Kombinationstyp:
–AA (6862, 62.1%), +AA (4184, 37.9%) ; –CC (778, 68.2%), +CC (363, 31.8%) ; –BB (546, 73.3%), +BB (199, 26.7%) ; –AB (742, 74.7%), +AB (251, 25.3%) ; –BA (464, 75.3%), +BA (152, 24.7%) ; –BC (373, 76.3%), +BC (116, 23.7%) ; –AC (615, 78.1%), +AC (172, 21.9%) ; –CA (253, 81.1%), +CA (59, 18.9%) ; –CB (223, 81.4%), +CB (51, 18.6%)

in %	Anzahl	Typ n	Typ $n+1$
NEGRA			
72.6	4639	A	A
6.2	396	A	B
4.9	313	C	C
4.3	275	A	C
3.5	225	B	B
3.5	221	B	A
2.1	135	B	C
1.8	112	C	A
1.1	71	C	B
TIGER			
68.4	7550	A	A
7.1	778	C	C
5.8	643	A	B
4.6	505	B	B
4.6	502	A	C
3.3	359	B	A
3.0	327	B	C
1.7	185	C	B
1.7	183	C	A
CZ			
75.9	1133	A	A
6.2	93	C	C
3.9	58	A	B
3.7	55	A	C
3.1	46	B	A
2.0	30	B	B
1.9	29	B	C
1.7	26	C	A
1.5	22	C	B

Tabelle 5.2: Verteilung der paarweisen Kombinationstypen A, B und C bei CNP in NEGRA, TIGER und CZ zwischen adjazenten Konjunktpaaren mit demselben Mutterknoten. Wenn mehr als 2 Konjunkte vorhanden sind in einer CNP, werden die inneren Köpfe zwei Mal berücksichtigt.

Typ	Anzahl	in %	± GN-Hauptkat.	Anzahl	in %
AA	11046	67.3	-	6862	41.8
			+	4184	25.5
CC	1141	7.0	-	778	4.7
			+	363	2.2
AB	993	6.1	-	742	4.5
			+	251	1.5
AC	787	4.8	-	615	3.7
			+	172	1.0
BB	745	4.5	-	546	3.3
			+	199	1.2
BA	616	3.8	-	464	2.8
			+	152	0.9
BC	489	3.0	-	373	2.3
			+	116	0.7
CA	312	1.9	-	253	1.5
			+	59	0.4
CB	274	1.7	-	223	1.4
			+	51	0.3

Tabelle 5.3: Verhältnis der Kombinationstypen zur Abdeckung mit Hauptkategorien in GermaNet über NEGRA, TIGER, CZ. Die Spalte „ ± GN-Hauptkat.“ drückt aus, ob die Köpfe eine gemeinsame Hauptkategorie in GermaNet besitzen.

5.2.3.1 Evaluation des Kombinationstyps +AA auf semantische Nähe

Wie semantisch nah sind die nominalen Köpfe, welche Kombinationstypen AA sind und mindestens eine GermaNet Hauptkategorie gemeinsam haben? Dazu wurden zufällig 100 Paare ausgewählt und intellektuell daraufhin evaluiert, ob die aus GermaNet berechnete gemeinsame Hauptkategorie eine semantische Nähe ausdrückt.

In den 100 Exemplaren befinden sich 2 Paare mit identischen Köpfen (in der Auswertung (221) markiert als „=“). Ein Fehler ergibt sich aus der Lemmatisierung durch GERTWOL (markiert als „x“), ein weiterer ist durch einen Schreibfehler („Zauber“ statt „Zauberer“) im Text selbst begründet (markiert als „t“).

(221) „+“ (91), „-“ (5), „=“ (2), „x“ (1), „t“ (1)

Die hohe Übereinstimmung zeigt, dass die Kombination der Hauptkategorie aus GermaNet zusammen mit der syntaktischen Kontext-Information eine zuverlässige semantische Klassifikation ergibt.

5.2.3.2 GermaNet-Kohyponymie in Konjunktköpfen

Neben der eher grobkörnigen Übereinstimmung in einer Hauptkategorie von GermaNet soll geklärt werden, wie oft Kohyponymie zwischen Konjunktköpfen besteht. Da jedes Synonympaar automatisch auch ein Kohyponympaar bildet, wurden die Synonyme ausgeblendet aus den Auswertungen.

Die Tabelle 5.4 auf der nächsten Seite zeigt die häufigsten der total 620 Köpfe von adjazenten CNP-Konjunkten aus allen drei Korpora, für die in GermaNet Kohyponymrelationen gefunden werden.

Kohyponymie ist sehr empfindlich auf die vertikale Strukturierung der Bedeutungen. Die Verwendung von künstlichen Konzepten ist in der Tabelle 5.4 markiert durch „?“ nach dem Oberbegriff. Die Feinheiten der Klassifikation können sich auf die Ausbeute der Kohyponymie negativ auswirken. So ist zum Beispiel bei der Repräsentation der Städte der politische Aspekt, ob eine Stadt Hauptstadt eines Staates oder Landes ist, in die Hyperonymie-Relation der Ortschaften einkodiert. Deshalb steht etwa „Köln“ und „Düsseldorf“ nicht in einer Kohyponymiebeziehung, weil „Düsseldorf“ im Gegensatz zu „Köln“ eine Landeshauptstadt bezeichnet.

Verteilung der Kohyponymie auf die Konjunkttypen Die Tabelle 5.5 auf Seite 254 schlüsselt die erkannten Kohyponyme auf die Konjunkttypen auf. Es zeigt sich, dass 1/6 der Kopfpaare des Typs AA Kohyponyme darstellen, aber nur 1/34 der Koordonyme des Typs CC. Beim Typ BB liegt das Verhältnis mit 1/22 etwa in der Mitte.

5.2.3.3 Das Vertiefen von Koordonymsets

Da viele Koordonyme vom Typ AA nur aus Einmal- oder Zweimalvorkommen bestehen, lassen sich zum Koordonymset eines Lemmas mittels gezielter Abfragen

in %	Anzahl	Lemma n	Lemma $n+1$	Oberbegriff
2.1	13	CDU	FDP	Partei
1.9	12	Junge	Mädchen	Kind
1.9	12	Jugendlich	Kind	junger Mensch ?
1.8	11	Frau	Mann	Verwandter
1.8	11	Frau	Mann	Erwachsener
1.8	11	ARD	ZDF	Fernsehanstalt
1.5	9	Samstag	Sonntag	Wochenendtag
1.3	8	Frau	Kind	Verwandter
1.0	6	Rom	Sinto	Zigeuner
1.0	6	Kroate	Serbe	Europäer
1.0	6	CDU	SPD	Partei
0.8	5	Schweiz	Österreich	Land
0.8	5	Portugal	Spanien	Land
0.8	5	Ost	West	Himmelsrichtung
0.8	5	Mutter	Vater	Mitmensch
0.8	5	Musik	Tanz	Darstellende Kunst
0.8	5	FDP	SPD	Partei
0.6	4	Kroatien	Slowenien	Land
0.6	4	Frankreich	Italien	Land
0.6	4	Bekannt	Freund	Mitmensch
0.5	3	Norden	Süden	Himmelsrichtung
0.5	3	Nation	Staat	politisches System
0.5	3	Kanada	USA	Land
0.5	3	Hoffnung	Sorge	Gefühl
0.5	3	Gemüse	Obst	Grünzeug
0.5	3	Deutschland	Frankreich	Land
0.5	3	Berlin	Hamburg	Stadtstaat
0.5	3	April	Mai	Frühlingsmonat
0.5	3	Afrika	Asien	Kontinent
0.3	2	Tango	Walzer	Tanz
0.3	2	Schüler	Student	Lerner
0.3	2	Sachsen	Thüringen	Bundesland
0.3	2	Rußland	UdSSR	Land
0.3	2	Rheinland-Pfalz	Saarland	Bundesland
0.3	2	Punkt	System	Computerterm ?
0.3	2	Polen	Ungarn	Land
0.3	2	Ostern	Pfingsten	kirchlicher Feiertag ?
0.3	2	Ordnung	Recht	juristischer Text ?
0.3	2	Oktober	September	Herbstmonat
0.3	2	November	Oktober	Herbstmonat

Tabelle 5.4: Kohyponymie zwischen CNP-Konjunktköpfen in NEGRA, TIGER und CZ. Gezeigt werden Fälle mit mindestens 2 Vorkommen. Die gemeinsamen Oberbegriffe sind in der 5. Spalte angezeigt.

in %	Anzahl	Typ	Kohyponymie	kumulativ
57.5	9313	AA	–	58
9.6	1556	AA	+	67
6.8	1103	CC	–	74
6.0	969	AB	–	80
4.8	772	AC	–	85
4.4	708	BB	–	89
3.7	601	BA	–	93
2.9	475	BC	–	96
1.9	307	CA	–	98
1.7	268	CB	–	99
0.2	32	BB	+	100
0.2	31	CC	+	100
0.1	23	AB	+	100
0.1	15	AC	+	100
0.1	14	BA	+	100
0.1	13	BC	+	100
0.0	6	CB	+	100
0.0	4	CA	+	100

Tabelle 5.5: Hyponymie in GermaNet aufgeschlüsselt nach Konjunkt-Typen

über grossen Textkorpora vertiefende Kookkurrenzdaten erheben.

So wurden für das AA-Koordonymset von „Politik“ aus Beispiel (222), welches 36 Lemmata umfasst, die Daten in Tabelle 5.6 auf der nächsten Seite via Google-Suche in deutschsprachigen Web-Seiten erhoben[2].

(222) **36 Politik** 12:Wirtschaft 3:Gesellschaft 3:Geschichte 2:Ökonomie 2:Öffentlichkeit 2:Wissenschaft 2:Wetter 2:Unternehmen 2:Kultur 2:Geld 1:Welt 1:Scheit 1:Regierung 1:Publikum 1:Psychologie 1:Polizei 1:Person 1:Medium 1:Marktsignal 1:Liebe 1:Kunst 1:Kommunalpolitik 1:Kind 1:Justiz 1:Institution 1:Ideologie 1:Gesetz 1:Geschäftemache 1:Finanz 1:Diplomatie 1:Demokratie 1:Bevölkerung 1:Beruf 1:Arbeitgeberverband 1:Arbeitgeber 1:etc.

Die Häufigkeit in den Koordonymsets ist ein Indikator für die Grössenordnung der Kookkurrenzen. Es gibt allerdings gewisse Reihenfolgeeffekte, welche zeigen, dass die Vorkommen von Koordinationspaaren nicht symmetrisch sind. So etwa im Fall von „Welt und Politik“ vs. „Politik und Welt“. Aufgrund der Diskussionen um den Nutzen und die Zuverlässigkeit von Frequenzdaten aus Websuchmaschinen (Kilgarriff 2007) wurden die Anfragen 2009 nochmals erhoben, wobei sich teilweise tatsächlich starke Änderungen ergaben.

[2]In grösseren linguistisch aufbereiteten Korpora wie etwa dem „DWDS Kerncorpus Version 170605b“ (Geyken 2004) finden sich nicht für alle Paare Belege.

K-Freq.	Google 2005	Google 2009	Anfrage
1	22.800	2.840	„Welt und Politik“
1	12.400	13.500	„Regierung und Politik“
1	9.660	17.600	„Beruf und Politik“
1	5.090	393	„Politik und Welt“
1	923	23.300	„Polizei und Politik“
1	898	16.400	„Politik und Regierung“
1	743	18.200	„Politik und Polizei“
1	614	2.390	„Politik und Psychologie“
1	581	4.570	„Psychologie und Politik“
1	554	44.000	„Politik und Beruf“
1	198	1.110	„Politik und Publikum“
1	98	947	„Publikum und Politik“
2	763.000	229.000	„Politik und Kultur“
2	211.000	314.000	„Kultur und Politik“
2	12.600	56.000	„Geld und Politik“
2	9.670	40.300	„Politik und Geld“
12	2.240.000	2.980.00	„Politik und Wirtschaft“
12	1.020.000	801.000	„Wirtschaft und Politik“

Tabelle 5.6: Frequenzdaten aus der Web-Suche für ausgewählte AA-Kordonyme zu „Politik“. Legende: K-Freq. = Anzahl der Vorkommen der Kooordonympaare

5.3 Koordonymie bei Adjektiven

Wie in Abschnitt 4.3.3 auf Seite 232 gezeigt, ist die Abdeckung von GermaNet bezüglich Adjektiven über Zeitungstextkorpora wie NEGRA und TIGER recht tief. In diesem Abschnitt werden die Resultate von Experimenten diskutiert, welche lexikalische Ressourcen aus den Köpfen von koordinierten CAP-Konjunkten erschliessen.

5.3.1 Extraktion von CAP aus Zeitungstexten

Aus den Zeitungstexten der NZZ-Ausgaben von Mai 1994 wurde ein Korpus aus knapp 900000 Token gebildet und mit Chunkie verarbeitet. Aus den 2155 CAP wurden danach alle Köpfe daraus extrahiert, welche paarweise dieselbe Wortart (ADJA oder ADJD) aufwiesen.

Im Extraktionsmodus, welcher nur reine Wort-Konjunkte berücksichtigt[3], finden sich 1443 Paare. Nach der Lemmatisierung der Köpfe ergeben sich daraus 1441 Koordonymsets. Da relativ wenige Paare zur Verfügung stehen, ist klar, dass viele Koordonymmengen nur 1 Element enthalten wie die Auflistung der Kardinalität in (223) zeigt.

(223) Auflistung der Verteilung der Kardinalität der Koordonymmengen (in Klammern sind absolute und relative Häufigkeit angegeben):
1 (885, 61.4%), 2 (253, 17.6%), 3 (115, 8.0%), 4 (75, 5.2%), 5 (30, 2.1%), 6 (21, 1.5%), 8 (18, 1.2%), 7 (16, 1.1%), 9 (7, 0.5%), 11 (5, 0.3%), 12 (4, 0.3%), 10 (3, 0.2%), 13 (3, 0.2%), 18 (2, 0.1%), 16 (1, 0.1%), 24 (1, 0.1%), 25 (1, 0.1%), 30 (1, 0.1%)

In der Tabelle 5.7 auf der nächsten Seite sind die grössten Kooordonymmengen aufgeführt. Obwohl sich interessante Kandidaten nebst Fehlern wie bei „letzt“ darunter befinden, ist die Datenmenge deutlich zu klein und von vielen Einmalvorkommen beherrscht.

Um das Problem der kleinen Datengrundlage zu mildern, wurde für jedes Lemma die Menge der möglichen Wortformen bestimmt und jeweils für jede Wortform maximal 250 Sätze aus dem Korpus des Wortschatz-Leipzig-Projekts über deren Web-Dienste (Biemann u. a. 2004a) bezogen, was insgesamt zu 1451240 Sätzen führte. Nach der Verarbeitung durch Chunkie liessen sich 97260 Koordonympaare extrahieren, welche 16429 Koordonymmengen ergaben.

Für einige Einträge aus der Tabelle 5.7 auf der nächsten Seite sind im Folgenden die vergrösserten Koordonymsets aufgeführt. Die Angabe „315:wirtschaftlich“ bedeutet dabei, dass in 315 Fällen das Lemma „politisch“ und „wirtschaftlich“ zusammen Kopf in einer CAP waren:

[3] Die Evaluationsresultate von Chunkie für tiefere CAP im Abschnitt 3.2.3.1 auf Seite 171 rechtfertigen diese Einschränkung.

Kopf	Koordonymset (mit Häufigkeit)
politisch	13:wirtschaftlich 10:sozial 4:kulturell 3:humanitär 2:ökonomisch 2:persönlich 2:militärisch 2:finanziell 2:ethnisch 1:ästhetisch 1:zahlenmäßig 1:wissenschaftlich 1:vorhanden 1:verbandspolitisch 1:säkular 1:sprachlich 1:sachlich 1:realwirtschaftlich 1:rechtlich 1:künstlerisch 1:individuell 1:ideologisch 1:gesellschaftlich 1:gemeinkriminell 1:geistig 1:ethisch 1:diplomatisch 1:biographisch 1:beruflich 1:akademisch
sozial	10:politisch 8:wirtschaftlich 5:ökonomisch 4:ökologisch 1:öffentlich 1:verantwortungsvoll 1:umweltbewußt 1:staatspolitisch 1:regulativ 1:psychisch 1:privat 1:persönlich 1:moralisch 1:mitfühlend 1:menschlich 1:medizinisch 1:materiell 1:landschaftspflegerisch 1:kulturell 1:juristisch 1:intellektuell 1:individualen 1:emotional 1:einsichtig 1:administrativ
groß	7:klein 5:kräftig 5:ganz 1:zusätzlich 1:stark 1:sperrig 1:schwer 1:schlank 1:romantisch 1:operiert 1:neu 1:mager 1:leer 1:korpulent 1:komfortabel 1:imposant 1:hager 1:fettgedruckt 1:farbig 1:dritt 1:clowneske 1:aufwendig 1:artistisch 1:alt
gut	3:billig 2:sicher 1:zufriedenstellend 1:schön 1:schnell 1:recht 1:rasant 1:populär 1:ostmitteleuropäisch 1:offenbar 1:interessant 1:gestylt 1:frisch 1:fernöstlich 1:erfolgreich 1:edel 1:bewährt 1:ausgeglichen
kulturell	4:wirtschaftlich 4:politisch 1:wissenschaftlich 1:technisch 1:sportlich 1:sozial 1:religiös 1:mäzenatische 1:mental 1:labil 1:gesellschaftlich 1:gemeinnützig 1:ethnisch 1:diffus 1:didaktisch 1:bürgerkriegsähnlich 1:beruflich 1:ander
letzt	2:dritt 2:14. 1:übersetzt 1:tolerierbar 1:semiotisch 1:sechst 1:lebend 1:jung 1:herausgegeben 1:entscheidend 1:elft 1:abschließend 1:42. 1:38. 1:34. 1:11.
alt	5:neu 1:schwerwiegend 1:recht 1:modern 1:link 1:jung 1:gültig 1:groß 1:gesellig 1:extrovertiert 1:erhalten 1:charmant 1:arbeitslos
demokratisch	2:republikanisch 1:willkürlich 1:vielgestaltig 1:unangemessen 1:unabhängig 1:stabil 1:sozialpolitisch 1:pluralistisch 1:nichtrassistisch 1:mehrsprachig 1:kontinuierlich 1:friedlich 1:frei
ökologisch	4:sozial 2:wirtschaftlich 1:ökonomisch 1:verantwortungsvoll 1:verkehrstechnisch 1:umweltbewußt 1:tierschützerisch 1:strukturell 1:mitfühlend 1:herrschaftsfrei 1:gesellschaftlich 1:einsichtig 1:bäuerlich

Tabelle 5.7: Exzerpt der grössten CAP-Koordonymmengen über dem NZZ-Korpus

- **politisch**: 315:wirtschaftlich 142:ökonomisch 141:sozial 106:militärisch 91:gesellschaftlich 88:religiös 79:kulturell 55:rassisch 52:rechtlich 43:moralisch 36:diplomatisch 31:administrativ 28:finanziell 26:soziologisch 24:persönlich 24:kirchlich 24:geographisch 23:psychologisch 23:ideologisch 22:strategisch 22:ethisch 19:juristisch 18:organisatorisch 18:intellektuell 16:humanitär 15:wissenschaftlich 15:historisch 15:ethnisch 14:technisch 13:geistig 12:bürgerlich 11:medial 11:konjunkturell 10:menschlich 10:künstlerisch 9:ästhetisch 9:personell 9:kommerziell 8:zeitgeschichtlich 8:philosophisch 8:geschichtlich 7:polizeilich 7:mental 7:literarisch 7:emotional 6:ökologisch 6:wirtschaftspolitisch 6:staatlich 6:pädagogisch 6:kommunikativ 6:geschäftlich 5:zivil 5:verfassungsmäßig 5:sprachlich 5:seelisch 5:poetisch 5:langfristig 5:fachlich 5:charakterlich 5:beruflich [...]

- **sozial** 175:ökologisch 151:wirtschaftlich 141:politisch 92:kulturell 88:ökonomisch 27:psychisch 27:ethnisch 25:ethisch 18:gesundheitlich 17:kommunikativ 17:demokratisch 16:pflegerisch 15:seelisch 14:religiös 14:pädagogisch 12:rechtlich 12:psychologisch 11:individuell 11:gesellschaftlich 10:technisch 10:emotional 9:wissenschaftlich 9:sportlich 9:rassisch 8:intellektuell 8:geistig 8:fiskalisch 8:beruflich 7:fachlich 7:erzieherisch 7:erfolgreich 6:sprachlich 6:solidarisch 6:sexuell 6:regional 6:mental 6:materiell 6:kraftvoll 6:ideologisch 6:gerecht [...]

- **alt**: 142:neu 69:jung 52:gebrechlich 19:krank 15:pflegebedürftig 11:schwach 10:alleinstehend 9:arm 8:bekannt 6:reif 6:gut 6:behindert 5:groß [...]

- **ökologisch**: 175:sozial 94:ökonomisch 36:wirtschaftlich 16:konventionell 15:ethisch 12:energiesparend 11:kulturell 9:feministisch 7:sozialpolitisch 6:zivil 6:technisch 6:sanft 6:politisch 6:gesellschaftlich 6:demokratisch 5:umweltfreundlich 5:regional 5:pazifistisch [...]

Die so gewonnenen Koordonymmengen sind damit gross genug geworden, um darin eine Rangordnung nach semantischer Ähnlichkeit zu versuchen. Die Verwendung der Anzahl der gemeinsamen Vorkommen ist stark von gemeinsamen Kookkurrenzen abhängig. Bei häufigen Lemmata wie „politisch“ oder „alt“ aus den obigen Beispielen ergibt sich daraus eine klare Reihenfolge. Bei Lemmata mit vielen verschiedenen, aber wenig häufigen Koordonymen ist die semantische Ähnlichkeit jedoch schwierig zu gewichten auf Grund der reinen Kookkurrenzen.

So gibt es für das Lemma „ausführlich“ in Beispiel (224a) zwar über 20 Kookkurrenzen, aber die Häufigkeitsunterschiede bleiben minim. Ähnliche Verhältnisse liegen bei „herablassend“ in (224c) vor. Für „sozialwissenschaftlich“ in Beispiel (224b) gibt es sogar nur Einmalvorkommen.

(224) a. **ausführlich** 2:sorgfältig 2:lang 2:informativ 2:facettenreich 2:differenziert 1:übersichtlich 1:öffentlich 1:zusammenfassend 1:völlig 1:vollstän-

dig 1:verdienstvoll 1:streng 1:sachlich 1:rasch 1:qualifizierend 1:pragmatisch 1:plastisch 1:irrwitzig 1:gut 1:genau 1:freundlich 1:detailliert 1:breit 1:aufschlußreichste 1:aktuell 1:akribisch

b. **sozialwissenschaftlich** 1:zeitgeschichtlich 1:technisch 1:soziologisch 1:sozialpolitisch 1:musisch 1:mathematisch 1:literarisch 1:juristisch

c. **herablassend** 2:wohlwollend 2:kühl 2:arrogant 1:verletzend 1:selbstsicher 1:oberflächlich 1:gehässig 1:drohend

Ähnlichkeit von Koordonymiemengen Ein Ansatz, um die semantisch ähnlicheren Koordonyme zu bestimmen, kann verfolgt werden, indem die Koordonymiemengen von 2 Koordonymen geschnitten werden. Die Kardinalität der Schnittmenge lässt sich danach als Gewicht der semantischen Nähe interpretieren zwischen den beiden Koordonymen (vgl. Biemann u. a. (2004b)).

Für die obigen Beispiele (224) ergeben sich dann bei einem Schwellwert von mindestens 3 gemeinsamen Elementen, die Rangordnungen in (225). Ein Eintrag wie „13:gut (420)“ bedeutet dabei: Das Lemma „gut“ hat bezüglich der Koordonymiemenge von „ausführliche“ 13 gemeinsame Lemma in seiner Koordonymmenge, wobei die Koordonymmenge von „gut“ die Kardinalität 420 aufweist.

(225) a. **ausführlich** 13:gut (420), 9:genau (116), 7:aktuell (128), 6:informativ (154), 5:detailliert (58), 5:sorgfältig (67), 5:breit (134), 5:lang (261), 4:sachlich (171), 4:rasch (178), 3:differenziert (80), 3:streng (135)

b. **sozialwissenschaftlich** 4:soziologisch (88), 3:literarisch (50), 3:musisch (60), 3:juristisch (64), 3:technisch (151)

c. **herablassend** 3:arrogant (118) 3:verletzend (61)

Im Vergleich zu den Listen in (224) zeigt sich einerseits eine Filterung, welche für das Lemma „ausführlich“ Wörter wie „pragmatisch“, „freundlich“ oder „irrwitzig“ entfernt. Andrerseits ergibt sich eine differenziertere Rangfolge in der Ähnlichkeit, sodass „streng“ weniger ähnlich ist als „lang“ oder „genau“.

Trotzdem gibt es noch Unschönheiten: Lemmas mit grossen Koordonymiesets wie „gut“, welche eher unspezifisch koordiniert werden, erhalten tendenziell hohe Rangierungen. Andrerseits ist bei „herablassend“ oder „sozialwissenschaftlich“ die Rangfolge immer noch unspezifisch. Eine Möglichkeit, die Nähe von spezifischen Wörtern zu verstärken, besteht darin, die Kardinalität der Koordonymmenge miteinzuberechnen.

Wenn k die Kardinalität der gemeinsamen Koordonyme darstellt und m die Kardinalität des zu gewichtenden Partners, dann sei $s = k^2/m$ das Ähnlichkeitsmass. Für obige Beispiele ergeben sich dann folgende Reihenfolgen:

(226) a. **ausführlich** 0.698:genau (116), 0.431:detailliert (58), 0.402:gut (420), 0.383:aktuell (128), 0.373:sorgfältig (67), 0.234:informativ (154), 0.187:breit (134), 0.112:differenziert (80), 0.096:lang (261), 0.094:sachlich (171), 0.090:rasch (178), 0.067:streng (135)

b. **sozialwissenschaftlich** 0.182:soziologisch (88), 0.180:literarisch (50), 0.150:musisch (60), 0.141:juristisch (64), 0.060:technisch (151)

c. **herablassend** 0.148:verletzend (61), 0.076:arrogant (118)

Wie man in (226) sieht, tritt der gewünschte Stauchungseffekt für grosse Koordonymmengen ein. Für eine objektive Validierung der Resultate sind jedoch systematische Auswertungen gegenüber bestehenden Ressourcen notwendig. Für die Berechnung von semantischer Nähe zwischen Koordonymen bietet sich zudem Synentropie (*mutual information*) an, ein Mass, das von Church und Hanks (1990) für semantische Ähnlichkeit etabliert wurde.

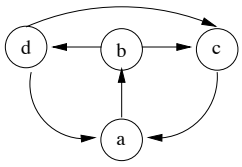
Kapitel 6

Baumbanken: Repräsentation, Suche und Extraktion

In diesem Kapitel werden die Grundlagen und Ansätze zur Repräsentation syntaktischer Strukturen präsentiert, welche sowohl für Baumbanken wie auch für die gespeicherten Resultate syntaktischer Analyseprogramme geeignet sind. Mit der Entwicklung von syntaktisch interpretierten Baumbanken, bei der wissenschaftshistorisch die Penn-Treebank für das Englische Marcus u. a. (1993) eine prägende Rolle gespielt hat, ist gleichzeitig das Bedürfnis entstanden, diese Ressourcen zu explorieren. Ein vorbildliche Fallstudie über den Nutzen von linguistisch interpretierten Korpora für Fragen der theoretisch ausgerichteten Linguistik hat Meurers (2005) für Deutsch gemacht. Während interaktive graphische Suchwerkzeuge wie TIGERSearch geeignet sind, um Beispiele und Gegenbeispiele schnell ausfindbar und durch die graphische Aufbereitung verifizierbar zu machen, eignet sich dieses Werkzeug schlechter für systematische Auswertungen syntaktischer Verhältnisse. Eine programmierbare Schnittstelle für die Suche, Extraktion und Transformation der Resultate ist für diese Zwecke unabdingbar.

Im Rahmen dieser Arbeit wurde deshalb eine Programmbibliothek in der logischen Programmiersprache PROLOG[1] entwickelt, welche ein einfaches, aber allgemeines und flexibles Repräsentationsformat und die dazugehörigen Primitive (Dominanz, lineare Präzedenz, Knotenbeschriftungen im Merkmal-Wert-Paar-Format, sekundäre Knotenverbindung) einer Abfragesprache für syntaktische Strukturen zur Verfügung stellt. Damit lassen sich von rein dependenzorientierten Formaten bis zu kombinierten phrasenstrukturell-funktionalen Formaten mit sekundärer Verlinkung wie bei TIGER alle gängigen Strukturbeschreibungen repräsentieren und über ein einheitliches Schnittstellen-Format absuchen.

[1] Eine gute deutschsprachige Einführung in diese Programmiersprache gibt Weisweber (1997). Eine auf den ISO-Prolog-Standard aktualisierte Version gibt Clocksin und Mellish (2003). Genau genommen handelt es sich dabei um zwei Standards 13211-1:1995 (1995); 13211-2:2000 (2000), wobei der letztere das Modul-System spezifiziert. Die Programmbibliothek sollte mit jedem ISO-PROLOG-kompatiblen PROLOG-System verwendbar sein.



$$G = \langle \{a,b,c,d\}, \{\langle a,b\rangle, \langle b,c\rangle, \langle b,d\rangle, \langle c,a\rangle, \langle d,a\rangle, \langle d,c\rangle\}\rangle$$

Abbildung 6.1: Gerichteter Graph (mit Zyklen) in graphischer und mengentheoretischer Darstellung

6.1 Graphen für syntaktische Strukturen

Die Repräsentation von endlichen syntaktischen Strukturen, welche mittels kontextfreier Grammatiken[2] (Hopcroft u. a. 2006) beschrieben werden können, ergibt sich in einfacher Weise aus graphentheoretischen Standard-Definitionen.

6.1.1 Definitionen und Terminologie

Gerichtete Graphen Ein gerichteter Graph $G = \langle N, E\rangle$ besteht aus einer endlichen, nicht-leeren Menge N von Knoten (*nodes*) und einer Menge E von Kanten (*edges*): $E \subseteq N \times N$. Eine Kante verbindet Knoten a mit Knoten b, genau dann wenn $\langle a,b\rangle \in E$. In einer Kante $\langle a,b\rangle$ heisst a Vorgänger von b und b Nachfolger von a.

Ein Pfad ist eine endliche Folge von Knoten, welche paarweise durch Kanten verbunden sind. Im Graphen von Abbildung 6.1 beispielsweise der Pfad $\langle d,c,a,b\rangle$. Die Länge eines Pfades ist die Anzahl der darin paarweise verbundenen Kanten, d.h. 1 weniger als die Anzahl seiner Knoten. Ein einfacher Pfad enthält einen Knoten höchstens einmal. Ein zyklischer Pfad ist ein einfacher Pfad, an dessen Ende sein Anfangsknoten nochmals angehängt wird. Ein Graph heisst zyklenfrei, wenn er nur einfache Pfade besitzt.

Gerichtete Bäume Ein gerichteter Baum ist ein zyklenfreier, gerichteter Graph mit folgenden Eigenschaften:

- Es gibt genau einen Knoten w, der keinen Vorgänger hat. Dieser Knoten heisst Wurzel.
- Jeder Knoten ausser der Wurzel hat genau einen Vorgänger.
- Von der Wurzel aus existiert genau ein Pfad zu jedem andern Knoten.

Die Höhe eines Baumes bezeichnet den längsten Pfad von der Wurzel aus.

[2] Komplexere Formalisierungen von syntaktischen Bäumen mit Merkmalstrukturen präsentiert Blackburn u. a. (1993). Carstensen u. a. (2004) enthält ein Einführung in die mengentheoretische Modellierung mit Blick auf computerlinguistische Strukturen.

Bei Bäumen nennt man in matrilinearer Sprechweise den Vorgänger gerne Mutter(knoten) und den Nachfolger Tochter(knoten). Knoten mit derselben Mutter heissen Schwester(knoten). Wenn man die Kanten eines Baums als unmittelbare Dominanzrelation (*immediate dominance*, *ID*) auffasst, dann dominiert ein Mutterknoten alle seine Tochterknoten unmittelbar.

Terminal-Knoten oder Blätter heissen Knoten, welche keinen Nachfolger (Tochter) besitzen. Nichtterminal-Knoten oder innere Knoten heissen Knoten, welche mindestens einen Nachfolger (Tochter) besitzen.

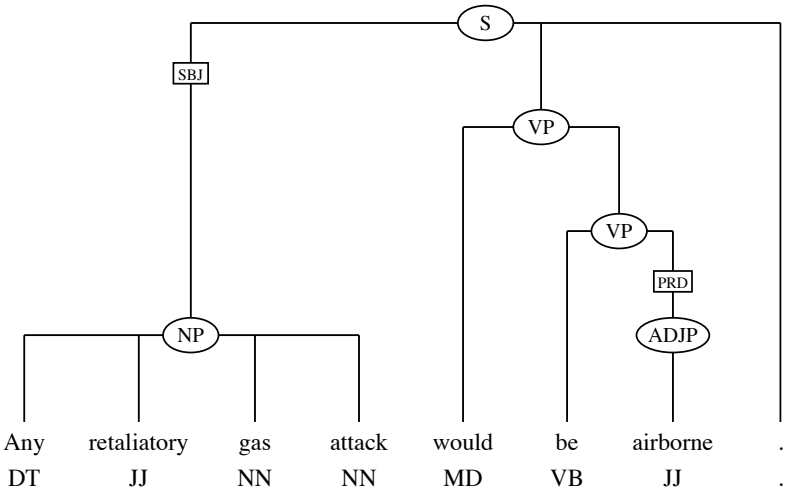
Geordnete Bäume Ein geordneter Baum ist ein Baum, bei dem zwischen allen Geschwistern die Knotenreihenfolge festgelegt ist. Diese Reihenfolge ist algebraisch formuliert eine Striktordnung zwischen allen Geschwisterknoten eines Baums, d.h. eine transitive und asymmetrische Relation. Diese Striktordnung heisst im Kontext von Syntaxbäumen lineare Präzedenz (*linear precedence*, *LP*). Sie kann als transitive Hülle der unmittelbaren linearen Präzedenzrelation (*immediate linear precedence*) aufgebaut werden, welche zwischen einem Knoten und seiner unmittelbar rechts davon stehenden Schwester besteht.

Markierte Bäume Bäume, welche für die Modellierung von syntaktischen Strukturen verwendet werden, haben normalerweise beschriftete, d.h. mit Merkmalswerten markierte Knoten. Bei den Blättern sind typischerweise die Werte der Merkmale Wortform und Wortart markiert, bei den inneren Knoten die syntaktische Kategorie und die Funktion. Eine partielle Knotenmarkierungsfunktion $m : N \rightarrow A$ bildet Knoten eines Baums aus der Menge N auf Attribute aus der Menge A ab. Manchmal werden auch Kanten markiert, d.h. ein Abbildung $m : E \rightarrow A$ zugeordnet. Ein markierter Baum ist ein Baum, dem mindestens eine Markierungsfunktion zugeordnet ist.

Syntaktische Bäume in Baumbanken Die Syntaxstrukturen der Penn-Treebank sind geordnete, markierte gerichtete Bäume, für die eine einfache Klammernotation ausreicht, wie sie in Abbildung 6.2 auf der nächsten Seite gezeigt wird. Sie haben alle ein Wurzel-Element und sämtliche Konstituenten und Terminale (auch Interpunktion) sind in die Struktur eingebaut. Partielle Markierungsfunktionen für die Blätter beschriften die Wortformen und Wortarten, sowie für die inneren Knoten die syntaktischen Kategorien. Für die syntaktischen Abhängigkeiten kann eine eigene Kanten-Markierungsfunktion verwendet werden, wie sie die TIGERSearch-Repräsentation in Abbildung 6.2 auf der nächsten Seite nahe legt. Genauso gut kann die Beschriftung der syntaktischen Abhängigkeit eines Tochterknotens bezüglich seiner Mutter als Knotenmarkierung der Tochter modelliert werden. Dies wird im kanonischen Format gemacht.

Die Syntaxstrukturen in NEGRA und TIGER erlauben Fragmente, d.h. mehrere Konstituenten, welche nicht zu einer einzelnen Wurzel zusammengefasst werden. Eine spezielle Form solcher Fragmente stellen die Interpunktionstoken dar,

```
( (S
    (NP-SBJ (DT Any) (JJ retaliatory) (NN gas) (NN attack) )
    (VP (MD would)
      (VP (VB be)
        (ADJP-PRD (JJ airborne) )))
    (. .) ))
```




```
% Unmittelbare Dominanz
id(84, 500, 0).
id(84, 500, 1).
id(84, 500, 2).
id(84, 500, 3).
id(84, 501, 6).
id(84, 502, 5).
id(84, 502, 501).
id(84, 503, 502).
id(84, 503, 4).
id(84, 504, 7).
id(84, 504, 500).
id(84, 504, 503).
id(84, 505, 504).

% Lineare Präzedenz
lp(84, 0, 1).
lp(84, 1, 2).
lp(84, 2, 3).
lp(84, 5, 501).
lp(84, 4, 502).
lp(84, 500, 503).
lp(84, 503, 7).
```

```
% Markierungsfunktionen
feature(84, 0, wd='Any').
feature(84, 0, cat='DT').
feature(84, 1, wd=retaliatory).
feature(84, 1, cat='JJ').
feature(84, 2, wd=gas).
feature(84, 2, cat='NN').
feature(84, 3, wd=attack).
feature(84, 3, cat='NN').
feature(84, 4, wd=would).
feature(84, 4, cat='MD').
feature(84, 5, wd=be).
feature(84, 5, cat='VB').
feature(84, 6, wd=airborne).
feature(84, 6, cat='JJ').
feature(84, 7, wd='.').
feature(84, 7, cat='.').
feature(84, 500, cat='NP').
feature(84, 500, fun='SBJ').
feature(84, 501, cat='ADJP').
feature(84, 501, fun='PRD').
feature(84, 500, cat='NP').
feature(84, 500, fun='SBJ').
feature(84, 502, cat='VP').
feature(84, 503, cat='VP').
feature(84, 504, cat='S').
feature(84, 505, cat='TOP').
```

Abbildung 6.2: Markierter gerichteter Syntaxbaum aus der Penn-Treebank (Satz 84 vom Brown-Korpus) in Klammernotation, in TIGERSearch-Darstellung und im PROLOG-Format

welche als isolierte Terminal-Knoten ohne jede Verbindung erscheinen.

Gerichteter Wald Ein azyklischer gerichteter Graph, bei dem jeder Knoten höchstens einen Vorgänger hat, heisst gerichteter Wald. Es kann in einem Wald also einzelne Knoten oder auch „Teilgraphen“ geben, welche nicht miteinander verbunden sind, und somit fehlt in solchen Fällen ein Wurzel-Knoten, von dem es zu jedem andern Knoten ein Pfad gibt.

Jeder gerichtete Wald kann in einen gerichteten Baum transformiert werden, indem ein neuer „künstlicher“ Wurzelknoten beigefügt wird, der zum Vorgänger aller bislang vorgängerlosen Knoten im Wald gemacht wird. Diese Technik wird im TIGERRegistry-Werkzeug zur Aufbereitung von Korpora für TIGERSearch verwendet, wo ein virtueller Wurzelknoten (VROOT) mit virtuellen Wurzelkanten eingefügt wird bei Syntaxstrukturen, welche keinen Wurzelknoten aufweisen.

Überkreuzende Kanten Um gerichtete Wälder (oder gerichtete Bäume als Spezialfall gerichteter Wälder) mit überkreuzenden Kanten[3] repräsentieren zu können, wie sie in syntaktischen Strukturen erscheinen, reicht eine zusätzliche Striktordnung aus, welche die lineare Präzedenz zwischen allen Blättern des Graphen ausdrückt.

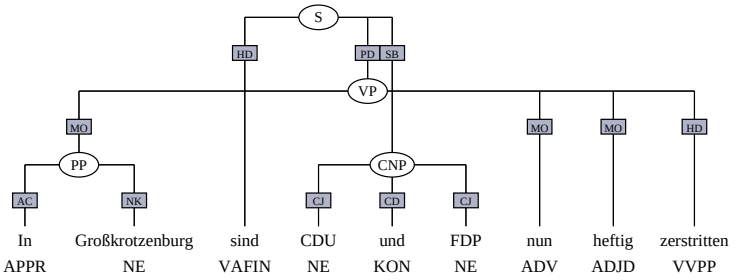
Sekundäre Kanten Sekundäre Kanten S eines gerichteten Graphen $G = \langle N, E \rangle$ sind wie die normalen Kanten $E \subseteq N \times N$ eine Teilmenge des Kreuzproduktes über den Knoten: $S \subseteq N \times N$. Es gibt dabei aber keine weiteren strukturellen Beschränkungen wie Zyklenfreiheit oder maximale Anzahl von sekundären Vorgängern. Die übliche Beschriftung von sekundären Kanten stellt formal eine Kantenmarkierung $m : S \rightarrow A$ dar. Eine Markierung an den Knoten ist nicht möglich, weil sekundäre Kanten im Gegensatz zu den primären mehr als einen Nachfolgerknoten haben können.

Syntaxgraphen Im Folgenden verwende ich den Ausdruck „Syntaxgraph“ als Überbegriff, um markierte geordnete gerichtete Wälder oder Bäume zu bezeichnen inklusive allenfalls benötigter zusätzlicher Relationen der Striktordnung der Terminale (überkreuzende Kanten) und der sekundärer Kanten mit ihrer Markierungsfunktion.

6.2 PROLOG-basierte Repräsentation von Baumbanken

Die Programmiersprache PROLOG stellt mit ihrem prädikatenlogischen Kern eine deklarative Beschreibungssprache für Relationen zur Verfügung. Die Repräsentati-

[3] Im Kontext der Dependenzgrammatik wird oft von nicht-projektiven Abhängigkeiten gesprochen (Foth u. a. 2004, 92).




```
% Unmittelbare Dominanz
id(12691, 500, 0).
id(12691, 500, 1).
id(12691, 503, 2).
id(12691, 501, 3).
id(12691, 501, 4).
id(12691, 501, 5).
id(12691, 502, 6).
id(12691, 502, 7).
id(12691, 502, 8).
id(12691, 502, 500).
id(12691, 503, 501).
id(12691, 503, 502).

% Lineare Präzedenz
lp(12691, 0, 1).
lp(12691, 3, 4).
lp(12691, 4, 5).
lp(12691, 500, 6).
lp(12691, 6, 7).
lp(12691, 7, 8).
lp(12691, 502, 2).
lp(12691, 2, 501).
```

```
% Markierungsfunktionen
feature(12691, 0, wd='In').
feature(12691, 0, cat='APPR').
feature(12691, 0, fun='AC').
feature(12691, 1, wd='Großkrotzenburg').
feature(12691, 1, cat='NE').
feature(12691, 1, fun='NK').
feature(12691, 2, wd=sind).
feature(12691, 2, cat='VAFIN').
feature(12691, 2, fun='HD').
feature(12691, 3, wd='CDU').
feature(12691, 3, cat='NE').
feature(12691, 3, fun='CJ').
feature(12691, 4, wd=und).
feature(12691, 4, cat='KON').
feature(12691, 4, fun='CD').
feature(12691, 5, wd='FDP').
feature(12691, 5, cat='NE').
feature(12691, 5, fun='CJ').
feature(12691, 6, wd=nun).
feature(12691, 6, cat='ADV').
feature(12691, 6, fun='MO').
feature(12691, 7, wd=heftig).
feature(12691, 7, cat='ADJD').
feature(12691, 7, fun='MO').
feature(12691, 8, wd=zerstritten).
feature(12691, 8, cat='VVPP').
feature(12691, 8, fun='HD').
feature(12691, 500, cat='PP').
feature(12691, 500, fun='MO').
feature(12691, 501, cat='CNP').
feature(12691, 501, fun='SB').
feature(12691, 502, cat='VP').
feature(12691, 502, fun='PD').
feature(12691, 503, cat='S').
```

Abbildung 6.3: Markierter gerichteter Syntaxbaum mit überkreuzenden Kanten aus NEGRA in TIGERSearch-Darstellung und im PROLOG-Format.

on von markierten geordneten gerichteten Wäldern kann deshalb die mengentheoretische Formalisierung fast wörtlich umsetzen.

Im Folgenden wird das kanonische PROLOG-Format definiert, das verwendet wurde, um eine einheitliche Schnittstelle für die Suche, Extraktion und Transformation der manuell erstellten und automatisch durch Parser berechneten Baumbanken zur Verfügung zu stellen.

Als anschauliche Beispiele für die kanonische Repräsentation ist in Abbildung 6.2 auf Seite 264 ein kontextfreier Syntaxgraph und in Abbildung 6.3 auf der vorherigen Seite ein Syntaxgraph mit überkreuzenden Kanten dargestellt.

Da in einer Baumbank eine Folge von Syntaxgraphen vorhanden ist, müssen die einzelnen Graphen auseinandergehalten werden durch eine Satznummer.

Knoten Die Terminalknoten eines Satzes mit n Token werden durch Ganzzahlen aus dem Intervall[4] $0..n-1$ in der Reihenfolge ihres Auftretens identifiziert. Die arithmetische Relation $<$ definiert dadurch implizit über den Terminalknoten die Striktordnung der linearen Präzedenz zwischen allen Blättern, welche für die überkreuzenden Kanten benötigt wird.

Die m Nichtterminal-Knoten eines Satzes werden per Konvention[5] durch Ganzzahlen aus dem Intervall $500..(499+m)$ identifiziert.

Kanten, d.h. unmittelbare Dominanz Die Relation der unmittelbaren Dominanz, welche durch Kanten ausgedrückt wird, repräsentieren die Fakten des Prädikats `id/3` (*immediate dominance*):

`id(SENT, FROM, TO)` ist wahr, gdw. im Satz SENT eine Kante vom Knoten FROM zum Knoten TO existiert.

Lineare Präzedenz Die Striktordnung zwischen den Schwesterknoten ist die transitive Hülle der Relation der paarweisen, unmittelbaren linearen Präzedenz, welche durch Fakten des Prädikates `lp/3` (*immediate lineare precendence*) gegeben ist:

`lp(SENT, LEFT, RIGHT)` ist wahr, gdw. sich im Satz SENT der Knoten LEFT unmittelbar links vom Schwesterknoten RIGHT befindet. D.h. kein Tochterknoten der Mutter von LEFT und RIGHT dazwischen erscheint.

Die Abbildung 6.3 auf der vorherigen Seite zeigt, dass für die Repräsentation von überkreuzenden Kanten im PROLOG-Format keine expliziten Präzendenzfakten gebraucht werden, vorausgesetzt die Terminalknoten sind entsprechend ihrer Reihenfolge durchnummeriert.

[4] Die Notation $a..b$ bezeichnet die Menge $\{ x \mid a \leq x \wedge x \leq b \}$.

[5] Dies entspricht der Usanz, welche im textbasierten Austauschformat NEGRA3 kodiert ist.

Markierungsfunktionen für Knoten Die verschiedenen Markierungsfunktionen m_i bis m_n, welche die Knoten mit Werten aus den verschiedenen Attributsmengen A_i bis A_n markieren, werden durch Fakten des Prädikats `feature/3` repräsentiert:

`feature(SENT, NODE, ATTRIBUTE = VALUE)` ist wahr, gdw. im Satz SENT der Knoten NODE von der Markierungsfunktion ATTRIBUTE den Wert VALUE zugewiesen bekommt.

Standardmässig werden die folgenden Prolog-Atome als ATTRIBUTE für die üblichen Markierungsfunktionen verwendet:

`cat` Syntaktische Kategorie von Phrasen und Wortart von Terminalsymbolen

`wd` Wortform von Terminalsymbolen

`fun` Syntaktische Funktion von Phrasen und Terminalsymbolen

`morph` Morphologische Tags der Terminalsymbole

Sekundäre Kanten Da sekundäre Kanten ohne Kantenmarkierung keinen Sinn machen, wird die ganze Information, welche eine Kante ausmacht, in Fakten des Prädikats `snde/4` abgelegt.

`snde(SENT, FROM, TO, VALUE)` ist wahr, gdw. im Satz SENT eine sekundäre Kante vom Knoten FROM zum Knoten TO mit der Kantenmarkierung VALUE existiert.

6.2.1 Deklarativität und Inferenz

Die deklarative Repräsentation von Syntaxgraphen als PROLOG-Wissensbasis ergibt in Kombination mit dem Inferenz-Mechanismus des PROLOG-Interpreters ein Suchwerkzeug. Die Frage „An welchen Stellen im Korpus kommt das Wort ‘CDU’ vor?“ lässt sich direkt als PROLOG-Anfrage umsetzen:

```
?- feature(Sent,Node,wd='CDU').
Node = 3,
Sent = 12691 ? ;
no
```

Das Beweisverfahren berechnet alle Lösungen für diese Anfrage über der Korpus-Wissensbasis (in diesem Fall nur der Satz aus Abbildung 6.3 auf Seite 266) und instantiiert die PROLOG-Variablen `Sent` und `Node` mit den entsprechenden Werten.

Komplexe Anfrage wie „Welche (anderen) Wörter mit der Wortart ’NE’ kommen im selben Satz mit dem Wort ‘CDU’ vor?“ sieht folgendermassen aus und ergibt 2 Lösungen[6]:

[6]Ohne die Verwendung des eingebauten PROLOG-Prädikats `dif/2`, welches für 2 Variablen nur die Instantiierung zu unterschiedlichen Termen erlaubt, wird auch „CDU“ selbst als Lösung

```
?- feature(Sent,Node,wd='CDU'), feature(Sent,Node2,cat='NE'),
   dif(Node,Node2), feature(Sent,Node2,cat='NE'),
   feature(Sent,Node2,wd=Word).
Node = 3,
Sent = 12691,
Word = 'Großkrotzenburg',
Node2 = 1 ? ;
Node = 3,
Sent = 12691,
Word = 'FDP',
Node2 = 5 ? ;
no
```

Für das Formulieren von „echten“ Korpus-Abfragen wäre es sehr mühsam, wenn alle Anfragen unmittelbar aus den grundlegenden Prädikaten `id/3`, `lp/3` und `feature/3` zusammengesetzt werden müssen.

Die wichtigen strukturellen Relationen innerhalb von Syntaxgraphen lassen sich in PROLOG deklarativ in entsprechenden Prädikaten definieren. So lässt sich die Dominanz-Relation `dom/3`, d.h. „Knoten *x* dominiert Knoten *y* in einem Satz“ in wenigen Programmzeilen rekursiv aus der unmittelbaren Dominanz als deren transitive Hülle definieren:

```
dom(Sent, Ancestor, Successor):-
   id(Sent,Ancestor, Successor).

dom(Sent, Ancestor, Successor):-
   id(Sent,Ancestor, Daughter),
   dom(Sent, Daughter, Successor).
```

Für alle wichtigen Relationen und Funktionen, welche sich bei den Abfragesprachen von Suchwerkzeugen wie TIGERSearch finden lassen, werden deshalb die entsprechenden PROLOG-Prädikate definiert und in der PROLOG-Bibliothek `cantreelib` zur Verfügung gestellt. Die aussagenlogischen Operatoren `,/2` (UND) sowie „`;/2`“ (ODER) stellt der PROLOG-Interpreter in der gewohnten Bedeutung zur Verfügung.

Negation Der Negationsoperator von PROLOG, `\+/1`, implementiert nicht die klassische Negation, sondern „*negation as failure*“ (NAF): `\+ P` ist wahr, gdw. `P` nicht bewiesen werden kann. Nun gilt im Zusammenhang von Baumbanken die sogenannte „*closed world assumption*“, d.h. wir gehen davon aus, dass alles, was an Sachverhalten in der Welt der Syntaxgraphen einer Baumbank existiert, durch die entsprechenden Fakten repräsentiert ist. Unter dieser Annahme und falls das

berechnet.

Beweisziel `\+ P` zur Laufzeit keine Variablen enthält, entspricht NAF der klassischen Negation. Die Anfrage „Ist das erste Wort von Satz 12691 kein Eigenname?“ lautet:

```
?- \+(feature(12691,0,cat='NE')).
yes
?- \+(feature(12691,0,cat='APPR')).
no
```

Wenn die Anfrage „Welche Wörter im Satz 12691 gibt es, welche keine Eigennamen sind?“ wie folgt formuliert wird, kommt eine vermeintlich „falsche“ Antwort zurück, falls man der NAF die klassische Negations-Bedeutung unterstellt. Denn obwohl vorhin bewiesen wurde, dass das 1. Wort im Satz ein Eigenname ist, soll es plötzlich keine Eigennamen mehr darin geben.

```
?-  \+(feature(12691,Node,cat='NE')).
no
```

Ungebundene Variablen im Skopus eines `\+`-Operators sind implizit allquantifiziert (und nicht existenzquantifiziert) und obige Anfrage hat somit die Bedeutung: „Gilt für alle Knoten, dass sie im Satz 12691 nicht Eigennamen sind?“. Das ist selbstverständlich nicht der Fall. Die ursprüngliche Anfrage kann trotzdem gestellt werden im Stil der üblichen Logik-Programmierungstechnik „Generiere-Und-Teste“[7].

```
?-  feature(12691,Node,cat=CAT), \+ CAT == 'NE'.
CAT = 'APPR',
Node = 0 ? ;
CAT = 'VAFIN',
Node = 2 ? ;
CAT = 'KON',
Node = 4 ?
...
```

Wie im Abschnitt 6.2.2 auf der nächsten Seite dokumentiert, stellt die Bibliothek `canontreelib` geeignete Knoten-Generierungsprädikate wie `node/2` zur Verfügung. Im Gebiet der Logik-Programmierung existieren zwar mittlerweile praktische Systeme (Muñoz-Hernández und Moreno-Navarro 2004), welche erweiterte Formen der Negation wie etwa konstruktive Negation unterstützen. Die erstellte Bibliothek sollte jedoch in beliebigen ISO-standard-kompatiblen PROLOG-Systemen benutzbar sein.

Zur Laufzeit ungebundene Variablen in negierten Beweiszielen sind durchaus nützlich, wenn man sich bewusst ist, was sie bedeuten. So gibt es in der Abfragesprache TIGERSearch keine Möglichkeit, die Anfrage „Welche Verbalphrasen

[7] Normalerweise benutzt man statt der negierten Identität von Termen (`\+ CAT == 'NE'`) gleich die Nicht-Identität (`CAT \== 'NE'`).

dominieren keine Nominalphrase?“ zu formulieren, da alle Knotenbeschreibungen implizit existenzquantifiziert sind (König und Lezius 2001, 22). So bedeutet etwa die TIGER-Anfrage [cat=VP"] >* [cat=(!"NP")]"das Folgende: „Welche VP dominieren irgendeine Phrase, welche etwas anderes als eine Nominalphrase ist?“. Eine allquantifizierende negierte Anfrage lässt sich hingegen in PROLOG problemlos formulieren:

```
?- feature(S,N,cat='VP'),
   \+(( feature(S,N2,cat='NP'),dom(S,N,N2) )).
```

6.2.2 PROLOG-Bibliothek `canontreelib`

Die folgenden Prädikate sind grundsätzlich logisch und generieren somit nichtdeterministisch alle möglichen Lösungen, wenn beim Prädikatsaufruf die Argumente nicht instantiiert sind. Argumente, welche beim Aufruf instantiiert sein müssen, sind wie üblich mit einem + markiert.

Dominanzrelationen

`dom(SENT, NODE, SUCCESSOR)` ist wahr, gdw. im Satz SENT der Knoten NODE über eine oder mehrere unmittelbare Dominanzrelationen mit dem Knoten SUCCESSOR verbunden ist.

`dom(SENT, NODE, SUCCESSOR, +MIN, +MAX)` ist wahr, gdw. im Satz SENT der Knoten NODE im Minimum über MIN und maximal über MAX unmittelbare Dominanzrelationen mit dem Knoten SUCCESSOR verbunden ist.

`sibling(SENT, NODE, SIBLING)` ist wahr, gdw. im Satz SENT die beiden unterschiedlichen Knoten NODE und SIBLING denselben Mutterknoten haben.

`leftcorner(SENT, NODE, LC)` ist wahr, gdw. im Satz SENT der Knoten LC derjenige Terminalknoten ist, der vom Knoten NODE dominiert wird, sodass der Pfad von NODE nach LC immer über die am weitesten links stehenden Tochterknoten führt. Jeder Terminalknoten steht in Left-Corner-Relation zu sich selbst.

`rightcorner(SENT, NODE, RC)` ist wahr, gdw. im Satz SENT der Knoten RC derjenige Terminalknoten ist, der vom Knoten NODE dominiert wird, sodass der Pfad von NODE nach RC immer über die am weitesten rechts stehenden Tochterknoten führt. Jeder Terminalknoten steht in Right-Corner-Relation zu sich selbst.

Präzedenzrelationen Für die Präzedenzrelationen, welche über die unmittelbare Präzedenz hinausgehen, gibt es in den verschiedenen Suchwerkzeugen leicht unterschiedliche Auffassungen (vgl. Steiner und Kallmeyer (2002)). Hier wird unterschieden zwischen der transitiven Hülle der unmittelbaren Präzedenzrelation (`mlp/3`), welche nur zwischen Geschwisterknoten definiert ist, und der Left-Corner-Präzedenz (`prec/3`), welche sich für alle Knoten aus der Terminal-Präzedenz ihrer Left-Corner-Knoten ergibt.

`mlp(SENT, NODE, AFTER)` (*mediate linear precedence*) ist wahr, gdw. im Satz SENT der Knoten NODE dieselbe Mutter wie der Knoten AFTER hat und NODE weiter links steht. In TIGERSearch heisst diese Relation „sibling with precedence“.

`mlp(SENT, NODE, AFTER, +MIN, +MAX)` ist wahr, gdw. im Satz SENT der Knoten NODE dieselbe Mutter wie der Knoten AFTER hat und NODE in deiner Distanz von mindestens MIN (MIN $>$ 0) und maximal MAX links von AFTER steht. Zwei Knoten in unmittelbarer Präzedenz zueinander haben die Distanz 1.

`prec(SENT, BEFORE, NODE)` ist wahr, gdw. im Satz SENT der Left-Corner des Knotens BEFORE sich vor dem Left-Corner des Knotens NODE befindet. Beachte: Da alle Left-Corner Terminalknoten sind, ergibt sich deren Präzedenz aus der Tokenreihenfolge.

`prec(SENT, BEFORE, NODE, +MIN, +MAX)` ist wahr, gdw. im Satz SENT der Left-Corner des Knotens BEFORE sich vor dem Left-Corner des Knotens NODE befindet und die Distanz zwischen den beiden Left-Cornern mindestens MIN (MIN $>$ 0) und maximal MAX beträgt.

Knoteneigenschaften

`root(SENT, NODE)` ist wahr, gdw. im Satz SENT der Knoten NODE keinen Vorgänger hat. Falls ein Syntaxgraph keinen Baum darstellt, kann es mehrere Wurzeln geben.

`arity(SENT, NODE, N)` ist wahr, gdw. im Satz SENT der Knoten NODE genau N Nachfolger hat.

`arity(SENT, NODE, +MIN, +MAX)` ist wahr, gdw. im Satz SENT der Knoten NODE höchstens MIN und maximal MAX Nachfolger hat.

`tokenarity(SENT, NODE, N)` ist wahr, gdw. im Satz SENT der Knoten NODE genau N Terminalknoten dominiert.

`tokenarity(SENT, NODE, +MIN, +MAX)` ist wahr, gdw. im Satz SENT der Knoten NODE mindestens MIN und maximal MAX Terminalknoten dominiert.

`continuous(SENT, NODE)` ist wahr, gdw. im Satz SENT der Knoten NODE kontinuierlich ist, d.h. jeder Terminalknoten zwischen dem Left- und dem Right-Corner von NODE, der nicht von NODE dominiert wird, wird auch von keinem andern Knoten dominiert.

`discontinuous(SENT, NODE)` ist wahr, gdw. im Satz SENT der Knoten NODE nicht kontinuierlich ist.

`node(SENT, NODE)` ist wahr, gdw. im Satz SENT der Knoten NODE vorkommt. Dieses Prädikat dient insbesondere als Generator. Die Reihenfolge, in der die Knoten instantiiert werden, ist nicht spezifiziert.

`nonterminal(SENT, NODE)` ist wahr, gdw. im Satz SENT der Nichtterminalknoten NODE vorkommt. Dieses Prädikat dient insbesondere als Generator. Die Reihenfolge, in der die Knoten instantiiert werden, ist nicht spezifiziert.

`terminal(SENT, NODE)` ist wahr, gdw. im Satz SENT der Terminalknoten NODE vorkommt. Dieses Prädikat dient insbesondere als Generator. Die Reihenfolge, in der die Knoten instantiiert werden, entspricht der Reihenfolge im Satz.

Funktionen für Knotenlisten Abfragen lassen sich oft kürzer formulieren, wenn Prädikate verwendet werden, welche Listen von Knoten berechnen. In Kombination mit Listenprädikaten höherer Ordnung zur Abbildung (*map*) oder Filterung (*filter*) ergeben sich konzise Notationen.

`ancestors(SENT, NODE, NODELIST)` ist wahr, gdw. die Knotenliste NODELIST alle Knoten der Reihe nach enthält, wie sie im Satz SENT ausgehend (aber nicht einschliessend) vom Knoten NODE bis zu dessen Wurzel erscheinen.

`daughters(SENT, NODE, NODELIST)` ist wahr, gdw. die Knotenliste NODELIST alle Tochterknoten von NODE in der Reihenfolge ihrer Präzedenz enthält.

`terminals(SENT, NODE, NODELIST)` ist wahr, gdw. die Knotenliste NODELIST alle Terminalknoten des Satzes SENT in der Reihenfolge ihrer Terminalpräzedenz enthält, welche vom Knoten NODE dominiert werden.

`terminals(SENT, NODELIST)` ist wahr, gdw. die Knotenliste NODELIST alle Terminalknoten des Satzes SENT in der Reihenfolge ihrer Terminalpräzedenz enthält.

`nonterminals(SENT, NODELIST)` ist wahr, gdw. die Knotenliste NODELIST alle Nichtterminalknoten des Satzes SENT einmal enthält. Die Reihenfolge wird festgelegt über die $<$-Relation der Knotennummern.

`sisters(SENT, +NODE, NODELIST)` ist wahr, gdw. die Knotenliste NODELIST alle Geschwisterknoten des gegebenen Knotens NODE in der Reihenfolge ihrer Präzedenz enthält.

Benutzerdefinierte Prädikate Mit der Möglichkeit, eigene PROLOG-Prädikate aus den Bibliotheks-Primitiven und den eingebauten Prädikaten zu definieren[8], lassen sich häufig gebrauchte Konfigurationsbeschreibungen kompakt abstrahieren. Die Relation „Knoten x dominiert Knoten y, bei dem das Attribut a den Wert v aufweist, d.h. $m_a(y) = v$“ kann folgendermassen definiert werden:

```
dom_fea(S,N1,N2,ATTR2):- feature(S,N2,ATTR2), dom(S,N1,N2).
```

Darauf aufbauend ergibt sich das Prädikat für die Menge aller Knoten, welche keine Knoten mit demselben Attribut dominieren.

```
nonrec_dom_fea(S,N,ATTR):-
   dom_fea(S,N,N2,ATTR), \+ dom_fea(S,N2,_,ATTR).
```

Satzweise Verarbeitung im Stapelmodus Wenn mehrere Tausend an Syntaxgraphen gleichzeitig im Arbeitspeicher von PROLOG gehalten werden, und zudem darüber mittels PROLOG-Beweiser gesucht wird, können auch moderne PROLOG-Systeme diese Ansprüche an Speicherplatz nicht erfüllen[9]. Eine Möglichkeit dieses Problem zu umgehen, wäre die Verwendung eines Standard-Datenbanksystems für das Speichern der Fakten, sodass mit PROLOG-Systemen[10] abfragemässig darauf zugegriffen werden kann. Da diese Lösungen nicht genügend allgemein zur Verfügung stehen, wurde eine einfache Schnittstelle geschaffen, welche Korpora im sequentiellen kanonischen PROLOG-Format satzweise aus reinen Textdateien einliest und verarbeitet.

Für das sequentielle kanonische PROLOG-Format gelten folgende Bedingungen:

- Alle Klauseln der Prädikate `lp/3`, `id/3`, `feature/3` und `snde/4` eines Satzes folgen sich direkt in der Datei. Die Reihenfolge dieser Klauseln ist nicht relevant.
- Nach der letzten Klausel befindet sich für jeden Satz eine Klausel `eos(SENT)` (*end of sentence*), welche das Ende der Satzrepräsentation von SENT markiert.

[8] Damit kann man über die Abstraktionsmöglichkeit hinausgehen, welche die TIGERSearch-Sprache im Prinzip mit dem nicht-rekursiven Template-Mechanismus (König und Lezius 2001, 24) anbietet.

[9] Ähnlich Probleme haben auch andere Korpus-API wie etwa das TIGER-API `http://tigerapi.sourceforge.net/` für Java, welches über dem TIGER-XML-Format vergleichbare Funktionalität zur Verfügung stellt wie unsere PROLOG-Bibliothek.

[10] SICStus PROLOG `http://www.sics.se/sicstus` unterstützt die Anbindung an die „Berkeley DB“. Das ciao-PROLOG-System `http://clip.dia.fi.upm.es/Software/Ciao/` eine Schnittstelle zu MySQL.

`process_corpus(+FILE, :HOOK)` liest sequentiell für jeden Satz die Klauseln aus der Korpusdatei FILE ein und ruft iterativ nach jedem Satz das benutzerdefinierte Prädikat HOOK auf, von dem alle Lösungen berechnet werden. Danach werden die aktuellen Satzklauseln gelöscht und der nächste Satz eingelesen. Da Backtracking verwendet wird, müssen die Ergebnisse des Hook-Prädikats entweder in die interne Wissensbasis abgespeichert oder herausgeschrieben werden.

Repräsentation diskontinuierlicher Grammatikregeln In Korpora, wo alle Sätze aus kontextfreien Bäumen besteht, lässt sich jedes Vorkommen einer Grammatikregel mit Hilfe des folgenden Prädikats beschreiben:

```
grammar_rule(S, LHS, RHS) :-
   daughters(S, LHS, RHS).
```

Für Korpora mit Syntaxgraphen, welche überkreuzenden Kanten aufweisen, gibt diese Repräsentation keine Hinweise darauf, ob und wo Diskontinuierlichkeit vorkommt. Eine milde Form der Diskontinuierlichkeit stellen auch Knoten dar, welche wie die Interpunktion isoliert auf der Terminalebene stehen. Diese Form der Diskontinuierlichkeit kann optional ignoriert werden.

`grammar_rule(SENT, LHS, RHS, OPTION)` ist wahr, gdw. im Satz SENT dem Mutterknoten LHS die Tochterspezifikation in der Liste RHS entspricht. Die Tochterspezifikation ist im Prinzip die Liste aller Tochterknoten in der Präzedenzreihenfolge. Zusätzlich werden alle Unterbrüche jeder Tochter auf der Terminalebene als Terme der Form `ref(0)` eingefügt und Fortsetzungen der `i`-ten Tochter (gezählt ab 1) nach Unterbrechungen als Terme der Form `ref(i)` eingefügt. Die Liste `OPTION` enthält Paare der Form `ATTRIBUTE : VALUE`, welche diejenigen Terminale bestimmen, welche nicht als Unterbrechungen zählen. Typischerweise sind das für Korpora wie NEGRA die Interpunktionen, d.h. `OPTION` hat den Wert `[cat:'$(',cat:'$.']`.

Beispiel: Der Syntaxgraph in Abbildung 6.3 auf Seite 266 ergibt für den Knoten VP eine RHS `[500,ref(0),6,7,8]`, da das unterbrechende Material nicht einer Tochterkonstituente zugehört, wird erscheint nur `ref(0)`. Somit sind alle Tochterknoten selbst kontinuierlich. Die RHS des Satzknotens S ergibt die Liste `[502,2,501,ref(1)]`, welche ausdrückt, dass der VP-Knoten 502 diskontinuierlich ist und nach der Tochterkonstituente 2 und 501 fortgesetzt wird.

Weiteres Die Bibliothek umfasst noch weitere Funktionalitäten, welche die Transformation in das kanonische Format und den Export in andere Korpus-Formate und Darstellungsformen (Kastendiagramme in LaTeXund HTML) unterstützen. Dies und weitere Informationen zu den Konversionswerkzeugen findet sich in der Dokumentation des Software-Pakets `canontree`.

Nicht direkt unterstützt wird die Verwendung von regulären Suchausdrücken in den Knotenbeschriftungen, obwohl für einzelne PROLOG-Systeme unterschiedliche Lösungen existieren[11], gibt es in der Logik-Programmierung noch keine systemübergreifende Spezifikation.

6.3 Andere Suchwerkzeuge

Im Folgenden werden andere Baumbank-Werkzeuge und ihre Unterschiede zur `canontree`-Software beleuchtet.

6.3.1 TIGERSearch

Das schon mehrfach erwähnte Werkzeug[12] ist ideal für den performanten, interaktiven Gebrauch, das Explorieren von Baumbanken und Zusammenstellen von einfachen statistischen Übersichten im integrierten Statistikwerkzeug. Zwar lassen sich Treffer von Suchanfragen exportieren, aber automatisierte Abfragen können nicht im Stapelverfahren gemacht werden.

Die Standard-Indizierung, welche mit dem separaten Werkzeug TIGERRegistry erfolgen muss, trennt Wortart (`pos`) und syntaktische Kategorie (`cat`) in zwei unterschiedliche Markierungfunktionen auf, da jedes Merkmal in TIGER entweder den Terminal- oder den Nichtterminalknoten zugeordnet werden muss. Für die Benutzung des integrierten Statistikwerkzeugs, das anhand von Merkmalen wie `pos` oder `cat` akkumuliert, ist dies für die Arbeit mit Korpora wie NEGRA oder TIGER letztlich ungünstig, da bedingt durch die flache Annotation Wörter und Phrasen oft dieselben Funktionen wahrnehmen. Ein weiterer Unterschied zur Repräsentation im kanonischen Format ist die Kodierung der Funktionslabel als Merkmal der unmittelbaren Dominanzrelation. Eine Anfrage, welche Knoten in der Funktion als Subjekt oder passiviertes Subjekt sucht, lautet

```
[ ] >SB #sb: [ ] | [ ] >SBP #sb: [ ]
```

Während in Knotenbeschreibungen innerhalb von eckigen Klammern sowohl die Verwendung von regulären Ausdrücken wie aussagenlogische Verknüpfungen konzise Notationen erlauben, stehen diese Mittel an etikettierten Dominanzrelationen nicht zur Verfügung. Ein Suchausdruck wie `[ fun=/SB.*/ ]` kann insbesondere deshalb nicht realisiert werden, da wie oben erwähnt, jedes Merkmal der Knotenbeschreibungen entweder auf Terminalknoten oder Nichtterminalknoten eingeschränkt ist.

[11] So hat G.V. Noord `http://www.let.rug.nl/~vannoord/prolog-rx` eine Anbindung für ISO-kompatible Reguläre Ausdrücke an SICStus Prolog zur Verfügung gestellt.

[12] Erhältlich unter `http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERSearch`.

6.3.2 Tgrep2

Für die in kontextfreier Klammerstruktur repräsentierten Syntaxbäume wie der Penn-Treebank steht mit dem textbasierten Befehlszeilenprogramm `tgrep2` (Rohde 2005) ein effizientes Werkzeug zur Verfügung, welches in Erweiterung zu seiner Vorgängerversion `tgrep` eine ausdrucksmächtige Suchsprache unterstützt, welche auch aussagenlogische Verknüpfungen von Teilstrukturbeschreibungen erlaubt. Ein kompakter Suchausdruck wie `NP !<< PP [> NP | >> VP]` lässt sich umschreiben mit „Alle NP, welche keine PP dominieren, aber selbst unmittelbar von einer NP oder mittelbar von einer VP dominiert sind.“ Dieselbe Suchanfrage definiert aus reinen Bibliotheksprädikaten[13]

```
np_pattern(S,NP):-
   feature(S,NP,cat='NP'),
   \+ ( feature(S,PP,cat='PP'), dom(S,NP,PP) ),
   ( feature(S,VP,cat='VP'), id(S,NP,VP)
   ; feature(S,NP2,cat='NP', dom(S,NP,NP2)
   ).
```

Als rein textbasiertes Befehlszeilen-Werkzeug unterstützt es Stapelverarbeitung ohne interaktives Explorieren der Resultate. Diese Sprache ist damit auch geeignet für web-basierte Abfragemöglichkeiten. Der Export und das Herausschreiben von Teilbäumen, welche von den Suchausdrücken gematcht werden, lässt sich durch Print-Operatoren in den Suchausdrücken steuern. Der Hauptnachteil von `tgrep` besteht darin, dass Syntaxgraphen mit überkreuzenden und sekundären Kanten nicht verarbeitet werden können. Diese Eigenschaft haben auch weitere Suchwerkzeuge aus dem englischsprachigen Raum wie „CorpusSearch“ (Randall 2000) oder „ICECUP“ (Wallis und Nelson 2000).

6.3.3 Weiteres

VICTORYA Eine Eigenheit dieses Suchwerkzeugs (Steiner und Kallmeyer 2002) ist, dass die Korpusdaten in einer SQL-Datenbank gespeichert und die Ausdrücke der Suchsprache in SQL-Anfragen übersetzt werden. Allerdings ist dieses Werkzeug auf Korpora mit rein kontextfreier Struktur eingeschränkt.

ANNIS [14] Ein web-basiertes Such- und Visualisierungswerkzeug, welches auch die Abfrage syntaktischen Relationen von Baumbanken unterstützt, ist in Dipper u. a. (2004) beschrieben. Im Gegensatz zu den bisher betrachteten Werkzeugen können satzübergreifende Anfragen gestellt werden, was im Kontext von Koreferenz-Annotation oft notwendig ist. Ein XML-basiertes Austausch-Format, das

[13] Wie in Abschnitt 6.2.2 auf Seite 274 gezeigt, lässt sich durch die Definition von Benutzerprädikaten ebenfalls eine kompakte Notation erreichen.

[14] Die Software und Dokumentation ist erhältlich unter `http://www.sfb632.uni-potsdam.de/~d1/annis/`.

Annotationen auf unterschiedlichen linguistischen Ebenen repräsentieren kann, ist in Dipper (2005) präsentiert.

netgraph Ein graphisches Suchwerkzeug in Client-Server-Architektur (Mírovský u. a. 2002) wurde im Rahmen der Prager Dependenz-Baumbank geschaffen. Überkreuzende Kanten können damit verarbeitet werden.

fsq Eine Abfragesprache, welche eine vollständige Logik 1. Stufe für endliche Strukturen (fsq ist ein Akronym für „finite structure query“) umfasst und sowohl überkreuzende als auch sekundäre Kanten behandeln kann, implementiert Kepser (2003) in Java. Obwohl eine geeignete Indizierung der Daten vorgenommen wird, kann aus Gründen der Berechnungskomplexität, welche der vollständigen Logik 1. Stufe prinzipiell innewohnt, die Beantwortung von Anfragen mit mehr als 3 ineinander verschachtelten Quantoren einige Minuten beanspruchen. Diese Implementation setzt somit das Konzept einer logik-basierten Abfragesprache sehr konsequent um, allerdings fehlen die für den praktischen Einsatz wichtigen Abstraktionsmöglichkeiten.

Kapitel 7

Schluss

In dieser Arbeit sind mit korpus- und computerlinguistischen Methoden verschiedene strukturelle Eigenschaften koordinierter Strukturen in der deutschen Sprache auf einer breiten empirischen Basis untersucht worden. Die Verbindung von automatischer qualitativer und quantitativer Beschreibung ist in dieser Form erst möglich geworden, seit syntaktisch konsistent interpretierte und annotierte Referenzkorpora im Umfang von Tausenden von Sätzen erstellt und verfügbar gemacht worden sind. Dies erlaubt eine Annäherung von beschreibender und quantitativer Linguistik (Volanovic und Köhler 2005).

Die Untersuchung hat dabei verschiedene Probleme des linguistischen Annotationsmodells von NEGRA/TIGER sowie Unzulänglichkeiten und Inkonsistenzen in der Annotation selbst zum Vorschein gebracht, welche im Zusammenhang mit Morphem-, Wort- oder Phrasenkoordinationen entstehen oder deren automatische Auswertung erschweren. So ist die Nichtannotation von nominalen Köpfen eine Schwäche dieser Ressourcen, da gerade enge Appositionen und komplexe Benennungen in schriftlichen Texten häufig auftreten.

Auch wenn mit diskontinuierlichen Strukturen und sekundären Kanten mächtige Beschreibungsmittel zur Verfügung stehen, welche nicht-lokale Strukturen fast unbeschränkt verknüpfen können, sollten nicht zuletzt im Hinblick auf die sprachtechnologische Verwendung solcher Ressourcen[1] die Strukturen möglichst lokal aufgebaut werden. Die Koordination von eng koordinierten Verbalköpfen ist dafür ein gutes Beispiel.

Ohne eine explizite Grammatik entsteht ein grosser Aufwand beim Explorieren der effektiv annotierten Verhältnisse. Ein provisorisches Annotationshandbuch reicht nicht aus, es braucht eine dokumentierte Grammatik, welche die vielfältigen Phänomene in realen Texten (inklusive der typisch schriftlichen Eigenheiten wie komplexe Benennungen, Zeitangaben usw.) systematisch beschreiben kann. Ausführlichere Annotationsrichtlinien sind zudem notwendig für konsistente Struktu-

[1] Plaehn (2004) hat gezeigt, dass eine direkte probablistische syntaktische Analyse von diskontinuierlichen Strukturen wie in NEGRA möglich ist, dass aber die Komplexität der Berechnungen dementsprechend steigt.

rierung.

Soweit im Rahmen dieser Arbeit sinnvoll und möglich wurde versucht, Bezüge zu Standardgrammatiken des Deutschen zu machen und Ansätze der theoretischen Linguistik beizuziehen. Da koordinierte Strukturen verbreitet als sekundäres Phänomen oder Grammatik zweiter Stufe betrachtet werden, ergibt sich für deren Behandlung eine starke Theorieabhängigkeit, welche für eine theorieneutrale Beschreibung, wie sie auch in den meisten Annotationsprojekten angestrebt wird, oft wenig abwirft. Die Untersuchung zur Verbreitung der Subjektlückenkonstruktion über dem TIGER-Korpus verbindet die beiden Welten etwas.

Die Evaluationen der verschiedenen Werkzeuge wie Chunker, dependenzorientierte Parsingverfahren und klassische Phrasenstrukturparser bezüglich Koordinationsstrukturen hat die spezifischen Probleme und Stärken dieser Systeme ausgeleuchtet. Die globalen Leistungsparameter weichen oft von Werten ab, welche man für koordinierte Strukturen noch erhält.

Um bewerten zu können, ob die erhobenen Merkmale bezüglich Häufigkeiten, Distanzeffekten, morphologischen und semantischen Ähnlichkeiten und Kommaklassifikation die Erkennung von koordinierten Strukturen unterstützen und optimieren können, müssen sie noch in ein Parsing-Modell eingebaut werden oder als nachgeschaltete *Ranking*-Komponente eingesetzt werden. Ersteres ist in Ansätzen wie dem in Abschnitt 3.5.2 auf Seite 197 vorgestellten Hamburger „papa“-Parser durch beliebig zuschaltbare gewichtete Beschränkungen möglich. Die Ansätze von Reranking (vgl. etwa Daumé III und Marcu (2004)) können an beliebige System angekoppelt werden, welche eine gute Vorauswahl von Lösungen liefern.

Die Erkennung der automatischen Klassifikation der Kommas mittels supervisierter Lernverfahren kann durch den Einbezug zusätzlicher Information sicher noch verbessert werden. Ob die Einberechnung von semantischer Nähe zwischen potentiellen Konjunkten nützlich ist oder nicht, ist mit der verwendeten Version von GermaNet nicht erschöpfend beantwortet. Experimente mit grösseren semantischen Ressourcen wären hier wünschenswert. Aus denselben Gründen mangelnder Abdeckung ist eine abschliessende Charakterisierung der Koordonymie in Bezug auf die traditionellen lexikalisch-semantischen Beziehungen wie Hyperonymie, Antonymie oder Meronymie nicht möglich.

Die Gewinnung von lexikalisch-semantischen Ressourcen aus Koordonymiemengen, welche aus den Resultaten syntaktischer Analysen destilliert werden, haben für Adjektive brauchbare Ergebnisse gebracht. Lexikalische und bedeutungsmässige Abdeckungsprobleme relationaler semantischer Ressourcen lassen sich korpusspezifisch vermindern. Die Ergebnisse müssen sicherlich mit kollokationsbasierten Ansätzen wie etwa Biemann und Osswald (2005) verglichen werden.

Die Programmbibliothek in der logischen Programmiersprache PROLOG für die automatische Auswertung der syntaktischen Beziehungen hat sich bei der Evaluation der Resultate verschiedener Parsing-Systeme und beim Untersuchen der strukturellen Eigenschaften der Syntaxgraphen in den Baumbanken bewährt. Die transparente und konzise Kodierung der grundlegenden grammatischen Beziehungen auf der Basis der unmittelbaren Dominanz, linearen Präzedenz sowie die Aus-

zeichnung der Struktur mit den jeweils verhandenen lexikalischen und syntaktischen Merkmalen liess sich auch für grosse Korpora mit dem Beweisverfahren von PROLOG direkt und einfach zur programmierbaren Abfrage und Transformation nutzen. XML-basierte Lösungen, welche bezüglich Standardisierung viele Vorteile bringen, erscheinen mir im Vergleich dazu immer noch schwerfälliger in der Handhabung.

Literaturverzeichnis

13211-1:1995 1995 13211-1:1995, ISO/IEC: *Information technology – Programming languages – Prolog – Part 1: General core*. ISO, Geneva, Switzerland, 1995

13211-2:2000 2000 13211-2:2000, ISO/IEC: *Information technology – Programming languages – Prolog – Part 2: Modules*. ISO, Geneva, Switzerland, 2000

Abeillé 2003 ABEILLÉ, Anne (Hrsg.): *Treebanks: Building and Using Parsed Corpora*. Bd. 20. Dordrecht : Kluwer Academic Publishers, 2003

Abney 1987 ABNEY, Steven: *The English Noun Phrase in its Sentential Aspect*, MIT, Cambridge, MA, Diss., 1987

Abney 1996 ABNEY, Steven: *Chunk Stylebook*. Working Draft. 1996

Albert u. a. 2003 ALBERT, Stephanie ; ANDERSSEN, Jan ; BADER, Regine ; BECKER, Stephanie: *TIGER Annotationsschema*. 2003

Altmann 1981 ALTMANN, Hans: *Formen der Herausstellung im Deutschen: Rechtsversetzung, Linksversetzung, Freies Thema und verwandte Konstruktionen*. Tübingen : Niemeyer, 1981 (Linguistische Arbeiten 105)

Bangalore und Joshi 1999 BANGALORE, Srinivas ; JOSHI, Aravind K.: Supertagging: An Approach to Almost Parsing. In: *Computational Linguistics* 25 (1999), Nr. 2, S. 237–265

Banko und Brill 2001 BANKO, Michele ; BRILL, Eric: Scaling to Very Very Large Corpora for Natural Language Disambiguation. In: *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, 2001, S. 26–33

Bartlett u. a. 2001 BARTLETT, II ; KOTRLIK, J. W. ; HIGGINS, C.: Organizational Research: Determining Appropriate Sample Size for Survey Research. In: *Information Technology, Learning, and Performance Journal* 19 (2001), Nr. 1, S. 43–50

Bayer 2002 BAYER, Joseph: Decomposing the Left Periphery: Dialectal and Cross-linguistic Evidence. In: FALK, Yehuda N. (Hrsg.): *Proceedings of Israel Association for Theoretical Linguistics* Bd. 18, 2002

Bergenholtz und Schaeder 1977 BERGENHOLTZ, Henning ; SCHAEDER, Burkhard: *Die Wortarten des Deutschen. Versuch einer syntaktisch orientierten Klassifikation*. Stuttgart : Klett, 1977

Berger u. a. 1996 BERGER, Adam L. ; PIETRA, Stephen A. D. ; PIETRA, Vincent D.: A Maximum Entropy Approach to Natural Language Processing. In: *Computational Linguistics* 22 (1996), Nr. 1, S. 39–71

Biemann u. a. 2004a BIEMANN, C. ; BORDAG, S. ; QUASTHOFF, U. ; WOLFF, C.: Web Services for Language Resources and Language Technology Applications. In: *Proceedings Fourth International Conference on Language Resources and Evaluation*, 2004

Biemann und Osswald 2005 BIEMANN, Chris ; OSSWALD, Rainer: Automatische Erweiterung eines semantikbasierten Lexikons durch Bootstrapping auf großen Korpora. In: FISSENI, B. (Hrsg.) ; SCHMITZ, H.-C. (Hrsg.) ; SCHRÖDER, B. (Hrsg.): *Sprachtechnologie, mobile Kommunikation und linguistische Ressourcen*, Peter Lang, 2005

Biemann u. a. 2004b BIEMANN, Christian ; BORDAG, Stefan ; QUASTHOFF, Uwe: Lernen paradigmatischer Relationen auf iterierten Kollokationen. In: *LDV-Forum* 19 (2004), Nr. 1/2, S. 103–111. – URL `http://ariadne.coli.uni-bielefeld.de/gldv/site/2004_Doppelheft/LDV-Forum2004.pdf`. – ISSN 0175-1336

Bies u. a. 1995 BIES, Ann ; FERGUSON, Mark ; KATZ, Karen ; MACINTYRE, Robert: *Bracketing Guidelines for Treebank II Style Penn Treebank Project 1*. 1995

Black u. a. 1991 BLACK, E. ; ABNEY, S. ; FLICKENGER, D. ; GDANIEC, C. ; GRISHMAN, R. ; HARRISON, P. ; HINDLE, D. ; INGRIA, R. ; JELINEK, F. ; KLAVANS, J. ; LIBERMAN, M. ; MARCUS, M. ; ROUKOS, S. ; SANTORINI, B. ; STRZALKOWSKI, T.: A Procedure for Quantitatively Comparing the Syntactic Coverage of English Grammars. In: *Speech and Natural Language: Proceedings of a Workshop Held at Pacific Grove, California, February 19-22, 1991*, Association for Computational Linguistics, 1991, S. 306–311

Blackburn u. a. 1993 BLACKBURN, P. ; GARDENT, C. ; MEYER-VIOL, W.: Talking About Trees. In: *Proceedings of the 6th Conference of the European Chapter of the Association for Computational Linguistics*, 1993, S. 21–29

Bohnet 2003 BOHNET, Bernd: Mapping Phrase Structures to Dependency Structures in the Case of Free Word Order Languages. In: *The First International Conference on Meaning-Text Theory*. Paris, 2003, S. 239–249

Bouma u. a. 2007 BOUMA, Gosse ; HENDRIKS, Petra ; HOEKSEMA, Jack: Focus particles inside prepositional phrases: A comparison of Dutch, English and German. In: *Journal of Comparative Germanic Linguistics* 10 (2007), Nr. 1, S. 1–24

Brants u. a. 2002 BRANTS, Sabine ; DIPPER, Stefanie ; HANSEN, Silvia ; LEZIUS, Wolfgang ; SMITH, George: The TIGER Treebank. In: *Proceedings of the Workshop on Treebanks and Linguistic Theories*. Sozopol, 2002. – URL `http://www.ims.uni-stuttgart.de/projekte/TIGER/paper/treeling2002.pdf`

Brants 1997 BRANTS, Thorsten: Internal and External Tagsets in Part-of-Speech Tagging. In: *Proceedings of Eurospeech*, 1997, S. 2787–2790

Brants 1999 BRANTS, Thorsten: *Tagging and Parsing with Cascaded Markov Models. Automation of Corpus Annotation*. Saarland University, 1999 (Saarbrücken Dissertations in Computational Linguistics and Language Technology 6)

Brants 2000a BRANTS, Thorsten: Inter-Annotator Agreement for a German Newspaper Corpus. In: *Second International Conference on Language Resources and Evaluation LREC-2000*. Athens, Greece, 2000

Brants 2000b BRANTS, Thorsten: TnT – A Statistical Part-of-Speech Tagger. In: *Proceedings of the Sixth Applied Natural Language Processing Conference ANLP-2000*, 2000, S. 224–231

Brants u. a. 2000 BRANTS, Thorsten ; DIPPER, Stefanie ; EISENBERG, Peter ; KRAMP, Sabine: *TIGER Annotationsschema*, 2000

Brants u. a. 1999 BRANTS, Thorsten ; HENDRIKS, Roland ; KRAMP, Sabine ; KRENN, Brigitte ; PREIS, Cordula ; SKUT, Wojciech ; USZKOREIT, Hans: *NEGRA Annotierschema*. 1999

Brants u. a. 2003 BRANTS, Thorsten ; SKUT, Wojciech ; USZKOREIT, Hans: *Treebanks: Building and Using Parsed Corpora*. Kap. Syntactic Annotation of a German Newspaper Corpus. Siehe (Abeillé 2003)

Bresnan 2001 BRESNAN, Joan: *Lexical-functional Syntax*. Malden, Mass. : Blackwell, 2001 (Blackwell Textbooks in Linguistics 16)

Büring und Hartmann 2001 BÜRING, Daniel ; HARTMANN, Katharina: The Syntax and Semantics of Focus-Sensitive Particles in German. In: *Natural Language & Linguistic Theory* 19 (2001), Nr. 2, S. 229–281

Buscha 1989 BUSCHA, Joachim: *Lexikon deutscher Konjunktionen*. Leipzig : Verlag Enzyklopädie, 1989

Butt u. a. 1999 BUTT, Miriam ; NIÑO, María-Eugenia ; SEGOND, Frédérique: *A Grammar Writer's Cookbook*. Stanford, CA : CSLI Publications, 1999

Camacho 2003 CAMACHO, José: *The Structure of Coordination*. Kluwer Academic Publishers, 2003

Carstensen u. a. 2004 CARSTENSEN, Kai-Uwe (Hrsg.) ; EBERT, Christian (Hrsg.) ; ENDRISS, Cornelia (Hrsg.) ; JEKAT, Susanne (Hrsg.) ; KLABUNDE, Ralf (Hrsg.) ; LANGER, Hagen (Hrsg.): *Computerlinguistik und Sprachtechnologie : Eine Einführung*. München : Elsevier, 2004

Chomsky 1957 CHOMSKY, Noam: *Syntactic Structures*. Den Haag : Mouton, 1957

Chomsky 1986 CHOMSKY, Noam: *Barriers*. 1986

Christ und Schulze 1995 CHRIST, Oli ; SCHULZE, B.M.: Ein flexibles und modulares Anfragesystem für Textcorpora. In: *Tagungsbericht des Arbeitstreffen Lexikon + Text*. Tübingen : Niemeyer, 1995

Church und Hanks 1990 CHURCH, Kenneth W. ; HANKS, Patrick: Word Association Norms, Mutual Information, and Lexicography. In: *Computational Linguistics* 16 (1990), Nr. 1, S. 22–29

Clematide 2002 CLEMATIDE, Simon: Selektive Evaluation von robusten Parsern. In: BUSEMANN, Stephan (Hrsg.): *Konvens 2002, 6. Konferenz zur Verarbeitung natürlicher Sprache, Proceedings*. Saarbrücken, September 2002, S. 23–29

Clocksin und Mellish 2003 CLOCKSIN, W.F. ; MELLISH, C.S: *Programming in Prolog*. 5. Auflage. Berlin : Springer, 2003

Daelemans und van den Bosch 2005 DAELEMANS, Walter ; BOSCH, Antal van den: *Memory-Based Language Processing*. Cambridge University Press, 2005

Dalheimer 1998 DALHEIMER, Matthias K.: *GNU-Tools zur Programmierung*. Köln : O'Reilly, 1998

Daum u. a. 2004 DAUM, Michael ; FOTH, Kilian ; MENZEL, Wolfgang: Automatic Transformation of Phrase Treebanks to Dependency Trees. In: *Proc. 4th Int. Conf. on Language Resources and Evaluation, LREC-2004*, 2004, S. 1149–1152

Daum u. a. 2003 DAUM, Michael ; FOTH, Kilian A. ; MENZEL, Wolfgang: Constraint Based Integration of Deep and Shallow Parsing Techniques. In: *Proceedings of EACL*, 2003, S. 99–106

Daumé III und Marcu 2004 DAUMÉ III, Hal ; MARCU, Daniel: NP Bracketing by Maximum Entropy Tagging and SVM Reranking. In: *Proceedings of EMNLP*. Barcelona, Spain, 2004

Dickinson 2005 DICKINSON, Markus: *Error Detection and Correction in Annotated Corpora*, The Ohio State University, Dissertation, 2005

Dipper 2003 DIPPER, Stefanie: *Implementing and Documenting Large-Scale Grammars — German LFG*, IMS, Universität Stuttgart, Dissertation, 2003

Dipper 2005 DIPPER, Stefanie: XML-based Stand-off Representation and Exploitation of Multi-Level Linguistic Annotation. In: ECKSTEIN, Rainer (Hrsg.) ; TOLKSDORF, Robert (Hrsg.): *Berliner XML Tage*, 2005, S. 39–50

Dipper u. a. 2004 DIPPER, Stefanie ; GÜTZE, Michael ; STEDE, Manfred ; WEGST, Tillmann: ANNIS: A Linguistic Database for Exploring Information Structure. In: *Interdisciplinary Studies on Information Structure (ISIS)* Bd. 1. Universitätsverlag Potsdam, 2004, S. 245–279

Drosdowski 2000 DROSDOWSKI, Günther (Hrsg.): *Duden Band 1: Die deutsche Rechtschreibung. Das umfassende Standardwerk auf der Grundlage der neuen amtlichen Regeln*. Mannheim/Wien/Zürich : Bibliographisches Institut, 2000

Dudenredaktion 2005 DUDENREDAKTION (Hrsg.): *Der Duden*. Bd. 4: *Duden, die Grammatik: Unentbehrlich für richtiges Deutsch*. 7. Dudenverlag, 2005

Eggs 2006 EGGS, Frederike: *Die Grammatik von als und wie*. Tübingen : Narr, 2006 (Tübinger Beiträge zur Linguistik 496)

Eisenberg 1999 EISENBERG, Peter: *Grundriß der deutschen Grammatik. Bd. 2: Der Satz*. Stuttgart : J.B. Metzler, 1999

Engelen 1978 ENGELEN, Bernhard: Zum Status des Elements *durch* in Sätzen wie *er ist durch den Wald durchgelaufen*. In: *Zeitschrift für germanistische Linguistik* 6 (1978), S. 178–186

Fortmann 2005 FORTMANN, Christian: Die Lücken im Bild von der Subjektlücken-Koordination. In: *Linguistische Berichte* 204 (2005), S. 441–476

Foth u. a. 2005 FOTH, Kilian ; MENZEL, Wolfgang ; SCHRÖDER, Ingo: Robust Parsing with Weighted Constraints. In: *Natural Language Engineering* 11 (2005), Nr. 1, S. 1–25

Foth 2004a FOTH, Kilian A.: *Eine umfassende Constraint-Dependenz-Grammatik des Deutschen*, 2004

Foth 2004b FOTH, Kilian A.: Writing Weighted Constraints for Large Dependency Grammars. In: *Recent Advances in Dependency Grammar, Workshop COLING 2004*, 2004

Foth u. a. 2004 FOTH, Kilian A. ; DAUM, Michael ; MENZEL, Wolfgang: A Broad-Coverage Parser for German Based on Defeasible Constraints. In: CHRISTIANSEN, Henning (Hrsg.) ; SKADHAUGE, Peter R. (Hrsg.) ; VILLADSEN, Jürgen (Hrsg.): *Constraint Solving and Language Processing*, 2004

Frank 2001 FRANK, A.: Treebank Conversion. Converting the NEGRA Treebank to an LTAG Grammar. In: *Proceedings of the Workshop on Multi-layer Corpus-based Analysis, Workshop of the EUROLAN*, 2001

Gaizauskas u. a. 1998 GAIZAUSKAS, R ; HEPPLE, M ; HUYCK, C: A Scheme for Comparative Evaluation of Diverse Parsing Systems. In: *Proceedings of the 1st International Conference on Language Resources and Evaluation (LREC'98)*. Granada, 1998, S. 143–149

Gallmann 1989 GALLMANN, Peter: *Syngrapheme an und in Wortformen. Bindestrich und Apostroph im Deutschen*. S. 85–110. In: EISENBERG, Peter (Hrsg.) ; HARTMUT, Günther (Hrsg.): *Schriftsystem und Orthographie*, Niemeyer, 1989 (Reihe Germanistische Linguistik 97)

Geyken 2004 GEYKEN, Alexander: Korpora als Korrektiv für einsprachige Wörterbücher. In: *Zeitschrift für Literaturwissenschaft und Linguistik* 136 (2004), S. 72–100

Grice 1975 GRICE, H.Paul: Logic and Conversation. In: COLE, P. (Hrsg.) ; MORGAN, J. (Hrsg.): *Syntax and Semantics: Speech Acts* Bd. 3. New York, 1975, S. 41–58

Haapalainen und Majorin 1994 HAAPALAINEN, Mariikka ; MAJORIN, Ari: *GERTWOL: Ein System zur automatischen Wortformenerkennung deutscher Wörter*. Helsinki : Lingsoft Oy, 1994

Haftka 1993 HAFTKA, Brigitta: Topologische Felder und Versetzungsphänomene. In: JACOBS, J. (Hrsg.) ; STECHOW, A. v. (Hrsg.) ; STERNEFELD, W. (Hrsg.) ; VENNEMANN, T. (Hrsg.): *Syntax: Ein internationales Handbuch zeitgenössischer Forschung*. Berlin : Walter de Gruyter, 1993

Hearst 1992 HEARST, Marti A.: Automatic Acquisition of Hyponyms from Large Text Corpora. In: *COLING 14: Proceedings of the 14th International Conference on Computational Linguistics* Bd. II, 1992, S. 539–545

Helbig und Buscha 1991 HELBIG, Gerhard ; BUSCHA, Joachim: *Deutsche Grammatik: Ein Handbuch für den Ausländerunterricht*. Leipzig : Langenscheidt, 1991

Höhle 1983 HÖHLE, Tilman: *Subjektlücken in Koordinationen*. Typoskript. 1983. – URL `http://www.uni-tuebingen.de/Deutsches-Seminar/hoehle/SLF-W5.1_neu.pdf`

Hopcroft u. a. 2006 HOPCROFT, John E. ; MOTWANI, Rajeev ; ULLMAN, Jeffrey D.: *Einführung in die Automatentheorie, Formale Sprachen und Komplexitätstheorie*. 2. Auflage. Pearson Studium, 2006

Houtman 1994 HOUTMAN, Joop: *Coordination and Constituency: A Study in Categorial Grammar*. Groningen Dissertations in Linguistics 13. 1994

IDS 2006 IDS: *GRAMMIS*. 2006. – URL `http://hypermedia.ids-mannheim.de/index.html`

Iwanska u. a. 2000 IWANSKA, Lucja ; MATA, Naveen ; KRUGER, Kellyn: *Fully Automatic Acquisition of Taxonomic Knowledge from Large Corpora of Texts: Limited-Syntax Knowledge Representation System based on Natural Language*. Kap. 10, S. 335–345. In: IWANSKA, L. (Hrsg.) ; SHAPIRO, S.C. (Hrsg.): *Natural Language Processing and Knowledge Processing*, MIT/AAAI Press, 2000

Jacobs 1983 JACOBS, Joachim: *Fokus und Skalen: Zur Syntax und Semantik der Gradpartikeln im Deutschen*. Tübingen : Niemeyer, 1983

Johannessen 1998 JOHANNESSEN, Janne B.: *Coordination*. New York : Oxford University Press, 1998 (Oxford Studies in Comparative Syntax)

Johnson 1998 JOHNSON, Mark: PCFG Models of Linguistic Tree Representations. In: *Computational Linguistics* 24 (1998), Nr. 4, S. 613–632

Karhiaho 2003 KARHIAHO, Izabela: *Der Doppelpunkt im Deutschen: Kontextbedingungen und Funktionen*. Göteborg : Acta Universitatis Gothoburgensis, 2003 (Göteborger germanistische Forschungen 42)

Kathol 2001 KATHOL, Andreas: Positional Effects in a Monostratal Grammar of German. In: *Journal of Linguistics* 37 (2001), S. 35–66

Kathol und Pollard 1995 KATHOL, Andreas ; POLLARD, Carl: On the Left Periphery of German Subordinate Clauses. In: *West Coast Conference on Formal Linguistics* Bd. 14, CSLI Publications/ SLA, 1995

Kepser 2003 KEPSER, Stephan: Finite Structure Query: A Tool for Querying Syntactically Annotated Corpora. In: *Proceedings of EACL*, 2003, S. 179–186

Kermes 2003 KERMES, Hannah: *Off-line (and On-line) Text Analysis for Computational Lexicography*, IMS, University of Stuttgart, Dissertation, 2003

Kilgarriff 2007 KILGARRIFF, Adam: Googleology is Bad Science. In: *Computational Linguistics* 33 (2007), Nr. 1, S. 147–151

Klein und Manning 2003 KLEIN, Dan ; MANNING, Christopher D.: Accurate Unlexicalized Parsing. In: HINRICHS, Erhard (Hrsg.) ; ROTH, Dan (Hrsg.): *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, 2003, S. 423–430

Klein 1993 KLEIN, Wolfgang: Ellipse. In: JACOBS, Joachim (Hrsg.) ; STECHOW, Arnim von (Hrsg.) ; WUNDERLICH, Dieter (Hrsg.): *Syntax. Ein internationales Handbuch zeitgenössischer Forschung* Bd. 1. Berlin/New York, 1993, S. 763–799

Klosa und Auberle 2001 KLOSA, Annette ; AUBERLE, Anette: *Duden Richtiges und gutes Deutsch: Wörterbuch der sprachlichen Zweifelsfälle*. Bd. 9. 5., neu bearb. Aufl., auf der Grundlage der neuen amtlichen Rechtschreibregeln. Mannheim : Dudenverlag, 2001

König und Lezius 2001 KÖNIG, Esther ; LEZIUS, Wolfgang: The TIGER Language. A Description Language for Syntax Graphs / IMS, University of Stuttgart. 2001. – Forschungsbericht

Kübler 2005 KÜBLER, Sandra: How Do Treebank Annotation Schemes Influence Parsing Results? Or How Not to Compare Apples And Oranges. In: *Proceedings of RANLP 2005*. Borovets, 2005

Kunze 2005 KUNZE, Claudia: *Semantik im Lexikon*. Kap. Semantische Relationstypen in GermaNet, S. 161–178, Gunter Narr Verlag, 2005

Lafferty u. a. 2001 LAFFERTY, John D. ; MCCALLUM, Andrew ; PEREIRA, Fernando C. N.: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In: *ICML*, 2001, S. 282–289

Lambrecht 1984 LAMBRECHT, Knud: Formulaicity, Frame Semantics, and Pragmatics in German Binomial Expressions. In: *Language* 60 (1984), Nr. 4, S. 753–796

Lee 1999 LEE, Nan-Hi: Zum Status der satzeinleitenden Komplementierer im Deutschen. In: *DOGILMUNHAK Koreanische Zeitschrift für Germanistik* 72 (1999), Nr. 4

Lezius u. a. 2000 LEZIUS, Wolfgang ; DIPPER, Stefanie ; FITSCHEN, Arne: IMSLex - Representing Morphological and Syntactical Information in a Relational Database. In: HEID, Ulrich (Hrsg.) ; EVERT, Stefan (Hrsg.) ; LEHMANN, Egbert (Hrsg.) ; ROHRER, Christian (Hrsg.): *Proceedings of the 9th EURALEX International Congress, Stuttgart, Germany*, 2000, S. 133–139

Lin 1995 LIN, Dekang: A Dependency-based Method for Evaluating-Broad Coverage Parsers. In: *Proceedings of IJCAI-95*, 1995

Lobin 1993 LOBIN, Henning: *Koordinationssyntax als prozedurales Phänomen*. Tübingen : Narr, 1993 (Studien zur deutschen Grammatik, 46)

Luschützky 2000 LUSCHÜTZKY, Hans C.: *Morphologie: ein internationales Handbuch zur Flexion und Wortbildung*. Kap. 46. Morphem, Morph und Allomorph, S. 451–462. In: BOOIJ, Geert (Hrsg.) ; LEHMANN, Christian (Hrsg.) ; MUGDAN, Joachim (Hrsg.): *Morphologie: ein internationales Handbuch zur Flexion und Wortbildung* Bd. Band 17. Berlin : Walter De Gruyter, 2000

Marcus u. a. 1993 MARCUS, Mitchell P. ; SANTORINI, Beatrice ; MARCINKIEWICZ, Mary A.: Building a Large Annotated Corpus of English: The Penn Treebank. In: *Computational Linguistics* 19 (1993), Nr. 2, S. 313–330

Meibauer u. a. 2002 MEIBAUER, Jörg ; DEMSKE, Ulrike ; GEILFUSS-WOLFGANG, Jochen: *Einführung in die germanistische Linguistik*. Stuttgart : Metzler, 2002

Meurers 2005 MEURERS, Walt D.: On the Use of Electronic Corpora for Theoretical Linguistics. Case studies from the Syntax of German. In: *Lingua* 115 (2005), Nr. 11, S. 1619–1639

Mírovský u. a. 2002 MÍROVSKÝ, Jiří ; ONDRUŠKA, Roman ; PRŮŠA, Daniel: Searching through Prague Dependency Treebank Conception and Architecture. In: *Proceedings of the Workshop on Treebanks and Linguistic Theories, September 20-21 (TLT02)*. Sozopol, Bulgaria, 2002

Müller 2004 MÜLLER, Frank H.: *Stylebook for the Tübingen Partially Parsed Corpus of Written German (TüPP-D/Z)*. 2004. – URL `http://www.sfs.uni-tuebingen.de/tupp/dz/stylebook.pdf`

Müller und Ule 2001 MÜLLER, Frank H. ; ULE, Tylman: Satzklammer annotieren und Tags korrigieren: Ein mehrstufiges Top-Down-Bottom-Up-System zur flachen, robusten Annotierung von Sätzen im Deutschen. In: LOBIN, Henning (Hrsg.): *Sprach- und Texttechnologie in digitalen Medien – Proceedings der GLDV-Frühjahrstagung 2001*, Norderstedt, 2001

Müller 1997 MÜLLER, Gereon: Beschränkungen für Binomialbildung im Deutschen. In: *Zeitschrift für Sprachwissenschaft* 16 (1997), Nr. 1, S. 5–51

Müller 2003 MÜLLER, Stefan: Mehrfache Vorfeldbesetzung. In: *Deutsche Sprache* 31 (2003), Nr. 1, S. 29–62

Muñoz-Hernández und Moreno-Navarro 2004 MUÑOZ-HERNÁNDEZ, Susana ; MORENO-NAVARRO, Juan J.: *Logic Programming*. Kap. Implementation Results in Classical Constructive Negation, S. 284–298, Springer, 2004

Noy und McGuinness 2001 NOY, Natalya F. ; MCGUINNESS, Deborah L.: Ontology Development 101: A Guide to Creating Your First Ontology / Stanford Knowledge Systems Laboratory. 2001 (SMI-2001-0880). – Forschungsbericht

van Oirsouw 1993 OIRSOUW, Robert R. van: Coordination (Koordination). In: JACOBS, Joachim (Hrsg.) ; STECHOW, Arnim von (Hrsg.) ; STERNEFELD, Wolfgang (Hrsg.) ; WUNDERLICH, Dieter (Hrsg.): *Syntax. Ein internationales Handbuch zeitgenössischer Forschung* Bd. 1. Berlin/New York : de Gruyter, 1993, S. 748–762

Olsen 1999 OLSEN, Susan: Durch den Park durch, zum Bahnhof hin. Komplexe Präpositionalphrasen mit einfachem direktionalem Kopf. In: WEGENER, Heide (Hrsg.): *Deutsch kontrastiv. Typologisch vergleichende Untersuchungen zur deutschen Grammatik*. Tübingen : Stauffenburg, 1999, S. 111–134

Pasch 2003 PASCH, Renate: *Handbuch der deutschen Konnektoren: Linguistische Grundlagen der Beschreibung und syntaktische Merkmale der deutschen Satzverknüpfer (Konjunktionen, Satzadverbien und Partikeln)*. Berlin : de Gruyter, 2003 (Schriften des Instituts für Deutsche Sprache Band 9)

Plaehn 2004 PLAEHN, Oliver: *Computing the Most Probable Parse for a Discontinuous Phrase Structure Grammar*. S. 91–106. In: BUNT, Harry (Hrsg.) ; CARROLL, John (Hrsg.) ; SATTA, Giorgio (Hrsg.): *New Developments in Parsing Technology*. Norwell, MA, USA : Kluwer Academic Publishers, 2004

Quinlan 1993 QUINLAN, J R.: *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993

Randall 2000 RANDALL, Beth: CorpusSearch User's Manual / University of Pennsylvania. 2000. – Technical Report

Ratnaparkhi 1998 RATNAPARKHI, Adwait: *Maximum Entropy Models for Natural Language Ambiguity Resolution*, University of Pennsylvania, Dissertation, 1998

Reis 1985 REIS, Marga: Satzeinleitende Strukturen im Deutschen. Über COMP, Haupt- und Nebensätze, w-Bewegung und die Doppelkopfanalyse. In: ABRAHAM, Werner (Hrsg.): *Erklärende Syntax des Deutschen*. Tübingen : Gunter Narr, 1985, S. 271–311

Reis 2005 REIS, Marga: On the Syntax of So-called Focus Particles in German - A Reply to Büring and Hartmann 2001. In: *Natural Language & Linguistic Theory* 23 (2005), S. 459–483

Reis und Rosengren 1997 REIS, Marga ; ROSENGREN, Inger: A Modular Approach to the Grammar of Additive Particles: the Case of German Auch. In: *Journal of Semantics* 14 (1997), Nr. 3, S. 237–309

Rizzi 1997 RIZZI, Luigi: The fine Structure of the Left Periphery. In: HAEGEMAN, L. (Hrsg.): *Elements of Grammar*. Dordrecht : Kluwer, 1997, S. 281–337

Rohde 2005 ROHDE, Douglas L. T.: *TGrep2 User Manual*, 2005. – URL `http://tedlab.mit.edu/~dr/Tgrep2/tgrep2.pdf`

Sampson 1995 SAMPSON, Geoffrey: *English for the Computer. The SUSANNE Corpus and Analytic Scheme*. Oxford : Clarendon Press, 1995

Schiehlen 2003 SCHIEHLEN, Michael: Combining Deep and Shallow Approaches in Parsing German. In: HINRICHS, Erhard (Hrsg.) ; ROTH, Dan (Hrsg.): *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*. Sapporo, Japan, 2003, S. 112–119

Schiehlen 2004 SCHIEHLEN, Michael: Annotation Strategies for Probabilistic Parsing in German. In: *Proceedings of the 20th International Conference on Computational Linguistics*. Geneva, August 2004

Schiller u. a. 1999 SCHILLER, Anne ; TEUFEL, Simone ; STÖCKERT, Christine: *Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und großes Tagset)*. 1999. – URL `http://www.ims.uni-stuttgart.de/projekte/corplex/TagSets/stts-1999.pdf`

Schmid 1995 SCHMID, Helmut: Improvements in Part-of-Speech Tagging with an Application to German. In: *Proceedings of the ACL SIGDAT-Workshop*. Dublin, 1995, S. 47–50

Schmid 2004 SCHMID, Helmut: Efficient Parsing of Highly Ambiguous Context-Free Grammars with Bit Vectors. In: *Proceedings of Coling 2004*, 2004, S. 162–168

Schwabe und Zhang 2000 SCHWABE, Kerstin (Hrsg.) ; ZHANG, Ning (Hrsg.): *Ellipsis in Conjunction*. Tübingen : Niemeyer, 2000 (Linguistische Arbeiten ; 418)

Skut 1999 SKUT, Wojciech: *Partial parsing for corpus annotation and text processing*. Saarbrücken : Saarland University, 1999 (Saarbrücken Dissertations in Computational Linguistics and Language Technology 10)

Skut 2001 SKUT, Wojciech: *Chunkie – A Statistical Partial Parser*, 2001

Skut u. a. 1997 SKUT, Wojech ; KRENN, Brigitte ; BRANTS, Thorsten ; USZKOREIT, Hans: An Annotation Scheme for Free Word Order Languages. In: *Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLP)*. Washington, D.C., 1997, S. 88–95

Stabler 1998 STABLER, Edward P.: Acquiring Languages with Movement. In: *Syntax* 1 (1998), Nr. 1, S. 72–97

von Stechow und Sternefeld 1988 STECHOW, Armin von ; STERNEFELD, Wolfgang: *Bausteine syntaktischen Wissens*. Opladen : Westdeutscher, 1988

Steedman 2002 STEEDMAN, Marc: *The Syntactic Process*. MIT Press, 2002

Steiner und Kallmeyer 2002 STEINER, Illona ; KALLMEYER, Laura: VIQTORYA–A Visual Query Tool for Syntactically Annotated Corpora. In: *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)*. Las Palmas, 2002, S. 1704–1711

Steiner 2003 STEINER, Petra: *Das revidierte Münsteraner Tagset / Deutsch (MT/D). Beschreibung, Anwendung, Beispiele und Problemfälle*. 2003. – URL `http://santana.uni-muenster.de/Publications/tagbeschr_final.ps`

Steiner 2004 STEINER, Petra: *Wortarten und Korpus: Automatische Wortartenklassifikation durch distributionelle und quantitative Verfahren*. Shaker Verlag, 2004

Sternefeld 2005 STERNEFELD, Wolfgang: *Syntax*. Stauffenburg, 2005

Teufel 1995 TEUFEL, Simone: A Support Tool for Tagset Mapping. In: *Proceedings of SIGDAT 1995. Workshop in cooperation with EACL 95*. Dublin, 1995

Thim-Mabrey 1988 THIM-MABREY, Christiane: Satzadverbialia und andere Ausdrücke im Vorvorfeld. In: *Deutsche Sprache* 1 (1988), S. 52–67

Tjong Kim Sang und Buchholz 2000 TJONG KIM SANG, Erik F. ; BUCHHOLZ, Sabine: Introduction to the CoNLL-2000 Shared Task: Chunking. In: CARDIE, Claire (Hrsg.) ; DAELEMANS, Walter (Hrsg.) ; NEDELLEC, Claire (Hrsg.) ; TJONG KIM SANG, Erik (Hrsg.): *Proceedings of CoNLL-2000 and LLL-2000*, Lisbon, Portugal, 2000, S. 127–132

te Velde 2005 VELDE, John R. te: *Deriving Coordinate Symmetries: A phase-based approach integrating Select, Merge, Copy and Match*. Amsterdam : John Benjamins, 2005

Volanovic und Köhler 2005 VOLANOVIC, Relja ; KÖHLER, Reinhard: *Syntactic Units and Structures*. Kap. 20. In: KÖHLER, R. (Hrsg.) ; ALTMANN, G. (Hrsg.) ; PIOTROWSKI, R. G. (Hrsg.): *Quantitative Linguistik. Ein internationales Handbuch. / Quantitative Linguistics*, de Gruyter, 2005

Volk 1999 VOLK, Martin: Choosing the Right Lemma When Analysing German Nouns. In: *Multilinguale Corpora: Codierung, Strukturierung, Analyse. 11. Jahrestagung der GLDV*. Frankfurt, 1999, S. 304–310

Volk 2001 VOLK, Martin: *The Automatic Resolution of Prepositional Phrase – Attachment Ambiguities in German*, Universität Zürich, Habilitationsschrift, 2001

Wallis und Nelson 2000 WALLIS, S ; NELSON, G: Exploiting Fuzzy Tree Fragment Queries in the Investigation of Parsed Corpora. In: *Literary and Linguistic Computing* 15 (2000), Nr. 3, S. 339–362

Wallis 2003 WALLIS, Sean: *Treebanks: Building and Using Parsed Corpora*. Bd. 20. Kap. Completing Parsed Corpora, S. 61–71. Siehe (Abeillé 2003)

Weisweber 1997 WEISWEBER, Wilhelm: *Prolog: Logische Programmierung in der Praxis*. Bonn : International Thomson Publishing, 1997

Zifonun u. a. 1997 ZIFONUN, Gisela ; HOFFMANN, Ludger ; STRECKER, Bruno: *Grammatik der deutschen Sprache*. Berlin; New York : de Gruyter, 1997 (Schriften des Instituts für deutsche Sprache; 7.1-3)

Zwischenstaatliche Kommission für Deutsche Rechtschreibung 2005 ZWISCHENSTAATLICHE KOMMISSION FÜR DEUTSCHE RECHTSCHREIBUNG (Hrsg.): *Deutsche Rechtschreibung: Regeln und Wörterverzeichnis: amtliche Regelung*. Narr, 2005

Anhang A

Tagsets

A.1 STTS-Wortartenkürzel

Kürzel	Kurz-Beschreibung
ADJA	Attributives Adjektiv
ADJD	Adverbiales oder prädikatives Adjektiv
ADV	Adverb
APPR	Präposition; Zirkumposition links
APPRART	Präposition mit Artikel
APPO	Postposition
APZR	Zirkumposition rechts
ART	Bestimmter oder unbestimmter Artikel
CARD	Kardinalzahl
FM	Fremdsprachliches Material
ITJ	Interjektion
KOUI	Unterordnende Konjunktion mit zu und Infinitiv
KOUS	Unterordnende Konjunktion mit Satz
KON	Nebenordnende Konjunktion
KOKOM	Vergleichspartikel, ohne Satz
NN	Normales Nomen
NE	Eigennamen
NNE	Kombination aus Nomen und Eigenname
PDS	Substituierendes Demonstrativpronomen
PDAT	Attribuierendes Demonstrativpronomen
PIS	Substituierendes Indefinitpronomen
PIAT	Attribuierendes Indefinitpronomen
PIDAT	Attribuierendes Indefinitpronomen mit Determiner
PPER	Irreflexives Personalpronomen
PPOSS	Substituierendes Possessivpronomen

Kürzel	Kurz-Beschreibung
PPOSAT	Attribuierendes Possessivpronomen
PRELS	Substituierendes Relativpronomen
PRELAT	Attribuierendes Relativpronomen
PRF	Reflexives Personalpronomen
PWS	Substituierendes Interrogativpronomen
PWAT	Attribuierendes Interrogativpronomen
PWAV	Adverbiales Interrogativ- oder Relativpronomen
PROAV	Pronominaladverb
PTKZU	zu vor Infinitiv
PTKNEG	Negationspartikel
PTKVZ	Abgetrennter Verbzusatz
PTKANT	Antwortpartikel
PTKA	Partikel bei Adjektiv oder Adverb
TRUNC	Kompositions-Erstglied
VVFIN	Finites Verb, voll
VVIMP	Imperativ, voll
VVINF	Infinitiv, voll
VVIZU	Infinitiv mit zu, voll
VVPP	Partizip Perfekt, voll
VAFIN	Finites Verb, aux
VAIMP	Imperativ, aux
VAINF	Infinitiv, aux
VAPP	Partizip Perfekt, aux
VMFIN	Finites Verb, modal
VMINF	Infinitiv, modal
VMPP	Partizip Perfekt, modal
XY	Nichtwort, Sonderzeichen
$,	Komma
$.	Satzbeendende Interpunktion
$(	Sonstige Satzzeichen; satzintern

A.2 Phrasale Kategorien

Kürzel	Englische Kurzbeschreibung	Kurz-Beschreibung
NP	noun phrase	Nominalphrase
AP	adjective phrase	Adjektivphrase
PP	adpositional phrase	Präpositionalphrase
S	sentence	Satz
VP	verb phrase (non-finite)	Infinite Verbalphrase
VZ	zu-marked infinitive	Infinitiv mit "zu"
CO	coordination	Koordination ungleicher Kategorien
AVP	adverbial phrase	Adverbphrase
AA	superlative phrase with "am"	Superlativphrase mit "am"
CNP	coordinated noun phrase	Koordinierte Nominalphrase
CAP	coordinated adjective phrase	Koordinierte Adjektivphrase
CPP	coordinated adpositional phrase	Koordinierte Präpositionalphrase
CS	coordinated sentence	Koordinierter Satz
CVP	coordinated verb phrase (non-finite)	Koordinierte Verbalphrase
CVZ	coordinated zu-marked infinitive	Koordinierter Infinitiv mit "zu"
CAVP	coordinated adverbial phrase	Koordinierte Adverbialphrase
MPN	multi-word proper noun	Eigenname aus mehreren Token
NM	multi-token number	Zahl aus mehreren Zifferntoken
CAC	coordinated adposition	Koordinierte Präposition
CH	chunk	Chunk
MTA	multi-token adjective	Adjektiv aus mehreren Token
CCP	coordinated complementiser	Koordinierter Komplementierer
DL	discourse level constituent	Konstituente auf Diskursebene
ISU	idiosyncratic unit	Idiosynkratische (nicht regelhafte) Einheit
QL	quasi-language	Quasi-Sprache

A.3 Grammatische Funktionen

Kürzel	Englische Kurz-Beschreibung	Kurz-Beschreibung
AC	adpositional case marker	Adposition
ADC	adjective component	Bestandteil eines Adjektivs
AG	genitive attribute	Genitiv-Attribut (TIGER)
AMS	measure argument of adj	Massangabe eine Adjektivs
APP	apposition	(lockere) Apposition
AVC	adverbial phrase component	Bestandteil einer Adverbialphrase
CC	comparative complement	Vergleichsergänzung
CD	coordinating conjunction	Koordinierende Konjunktion
CJ	conjunct	Konjunkt
CM	comparative conjunction	Vergleichskonjunktion
CP	complementizer	Komplementierer
CVC	collocational verb construct	Funktionsverbgefüge
DA	dative	Dativ-Objekt
DH	discourse-level head	Einleitungssatz für direkte Rede
DM	discourse marker	Gesprächspartikel
EP	expletive “es”	Expletives “es”
GL	prenominal genitive	Pränominaler Genitiv (NEGRA)
GR	postnominal genitive	Postnominaler Genitiv (NEGRA)
HD	head	Kopf
MNR	postnominal modifier	Postnominaler Modifikator
MO	modifier	Modifikator
MR	rhetorical modifier	Rhetorischer Modifikator
MW	way (directional modifier)	Richtungsangabe
NG	negation	Negation
NK	noun kernel modifier	Modifikator des Nominalkerns
NMC	numerical component	Bestandteil eines mehrteiligen nummerischen Ausdrucks
OA	accusative object	Akkusativ-Objekt
OA2	second accusative object	Zweites Akkusativ-Objekt
OC	clausal object	Objektsatz
OG	genitive object	Genitiv-Objekt
OP	prepositional object	Präpositional-Objekt
PAR	parenthesis	Parenthese
PD	predicate	Prädikativergänzung
PG	pseudo-genitive	Pseudo-Genitiv (“von”-Konstruktion)

Kürzel	Englische Kurz-Beschreibung	Kurz-Beschreibung
PH	placeholder	Platzhalterkonstruktion
PM	morphological particle	Morphologisches Partikel
PNC	proper noun component	Eigennamenbestandteil
RC	relative clause	Relativsatz
RE	repeated element	Wiederholtes Element
RS	reported speech	Indirekte Rede
SB	subject	Subjekt
SBP	passivised subject (PP)	Passiviertes Subjekt
SP	subject or predicate	Subjekt oder Prädikativergänzung
SVP	separable verb prefix	Abtrennbares Verbpräfix
UC	(idiosyncratic) unit component	Bestandteil einer unanalysierbaren Einheit
VO	vocative	Vokativ (Anredeform)

Lebenslauf

Name	Simon Remo Clematide
Adresse	Langmauerstr. 76 8006 Zürich siclemat@cl.uzh.ch
Geboren	23.3.1968 in Romanshorn TG
Heimatort	Amriswil TG
Zivilstand	Verheiratet mit Renate Clematide-Müller seit 1991
Kinder	Joachim (1991), Norma (1995), Paula (2001)

Ausbildung

1975 – 1983	Primar- und Sekundarschule in Romanshorn
1983 – 1987	Kantonsschule Romanshorn
1987	Eidg. Anerkannte Kantonale Maturität Typus B
1988 – 1990	Ausbildung zum Primarlehrer „Maturitätsgebundener Weg“ am Lehrerseminar Kreuzlingen
1990	Thurgauer Primarlehrer-Patent
1990 – 1999	Studium der Germanistik, Informatik und Philosophie an der Universität Zürich
1999	Lizentiatsabschluss
2000 – 2006	Doktoratsstudium in Computerlinguistik an der Universität Zürich

Berufstätigkeit

1990 – 1997	Vikariate auf Primarschulstufe
1999 – 2005	Assistenz am Institut für Computerlinguistik, Universität Zürich (Teilzeit)
2002 – 2009	Programmierung von Web- und Serverdiensten, buero m, Zürich (Teilzeit)
2006 – 2009	Wissenschaftlicher Mitarbeiter am Institut für Computerlinguistik (Teilzeit)